MODELAMIENTO PREDICTIVO DEL GANADOR DE UN PARTIDO DE FÚTBOL DE LA CATEGORÍA A DEL FÚTBOL PROFESIONAL COLOMBIANO USANDO APRENDIZAJE DE MÁQUINA

EDWIN FERNANDO ARIAS ROA

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA INGENIERÍA DE SISTEMAS
BOGOTÁ D.C.
2019

MODELAMIENTO PREDICTIVO DEL GANADOR DE UN PARTIDO DE FÚTBOL DE LA CATEGORÍA A DEL FÚTBOL PROFESIONAL COLOMBIANO USANDO APRENDIZAJE DE MÁQUINA

EDWIN FERNANDO ARIAS ROA

Trabajo de grado presentado como requisito parcial para optar al título de: ingeniero de sistemas

Asesor: Roger Enrique Guzmán
M. Sc. (c) Ingeniería de Sistemas y Computación.

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
BOGOTÁ D.C.
2019

Nota de Aceptación

Aprobado por el comité de grado en cumplimiento de los requisitos exigidos por la Facultad de Ingeniería y la Universidad Católica de Colombia para optar al título de ingeniero de sistemas.

_____
Director

_____
Jurado 1

_____
Jurado 2

_____
Revisor Metodológico

BOGOTÁ D.C, MAYO 30 DE 2019

AGRADECIMIENTOS

En primer lugar, quiero agradecer a Dios por todas las bendiciones recibidas a lo largo de mi vida, por guiarme en mente y espíritu hacia la realización de mis metas en los ámbitos profesional, laboral y profesional. Así mismo, quiero agradecer a mi familia por su inmenso apoyo en diferentes aspectos durante todo este proceso de formación académica universitaria.

De igual forma, quiero expresarle mis agradecimientos a la universidad por aceptarme y permitirme ser un egresado que pregone los valores del buen ciudadano, a cada uno de los profesores que compartieron sus conocimientos y experiencias para permitirme desarrollarme como profesional, a mis compañeros de carrera quiénes fueron clave fundamental en este proceso de formación, y especialmente a mi tutor de trabajo de grado Roger Guzman por el apoyo y acompañamiento incondicional siempre brindando su conocimiento y experiencia en pro del desarrollo de este proyecto.

CONTENIDO

Pág.

TABLA DE ILUSTRACIONES 5
TABLA DE TABLAS 7
LISTA DE ABREVIATURAS 8
PALABRAS CLAVE 9
GLOSARIO 10
RESUMEN 11
INTRODUCCIÓN 12
0. FICHA TÉCNICA 13
1. GENERALIDADES 14
1.1. ANTECEDENTES 14
1.2. PLANTEAMIENTO DEL PROBLEMA 15
1.2.1 Descripción del problema. 15
1.2.2 Formulación del problema. 16
1.3. OBJETIVOS 17
1.3.1 Objetivo General. 17
1.3.2 Objetivos específicos. 17
1.4. JUSTIFICACIÓN 18
1.5. ALCANCES Y LIMITACIONES 19
1.5.1 Alcance. 19
1.5.2 Limitaciones. 19
1.6. MARCO REFERENCIAL 20
1.6.1 Marco teórico. 20
1.6.1.1 Algoritmo Naïve Bayes. 20
1.6.1.2 Algoritmo Random Forest. 21
1.6.1.3 Máquinas de Soporte Vectorial. 22
1.6.1.4 Regresión lineal. 23
1.6.1.5 Regresión Logística Binaria. 25
1.6.1.6 Regresión Logística Multinomial. 27
1.6.1.7 K nearest neighbor. 28
1.6.1.8 Métricas de desempeño. 29

1.6.2 Marco Conceptual. 35
1.6.2.1 Fútbol. 35
1.6.2.2 Minería de datos. 36
1.6.2.3 Aprendizaje de máquina. 36
1.6.2.4 Aprendizaje profundo. 38
1.6.3 Marco Legal. 38
1.6.3.1 Habeas Data. 38
1.6.3.2 Aviso legal Diario AS. 39
1.7. ESTADO DEL ARTE 40
1.8. METODOLOGÍA 43
1.8.1 Conjunto de datos. 43
1.8.2 Preprocesamiento y limpieza de datos. 44
1.8.3 Identificación de características. 44
1.8.4 Muestreo. 45
1.8.5 Construcción modelo de predicción. 45
1.8.6 Salida. 45
2. DESARROLLO DE LA METODOLOGÍA Y RESULTADOS 47
2.1. CONJUNTO DE DATOS 47
2.2. PREPROCESAMIENTO Y LIMPIENZA DE INFORMACIÓN 50
2.3. IDENTIFICACIÓN Y SELECCIÓN DE CARACTERISTICAS 51
2.4. MUESTREO 62
2.5. CONSTRUCCIÓN MODELO PREDICTIVO 65
2.5.1 Regresión Logística multiclase. 65
2.5.2 Random Forest. 68
2.5.3 Máquina de soporte vectorial. 71
2.5.4 Puntuación algoritmos de aprendizaje de máquina. 74
3. CONCLUSIONES 76
4. RECOMENDACIONES Y TRABAJO FUTURO 77
4.1. RECOMENDACIONES 77
4.2. TRABAJO FUTURO 77
BIBLIOGRAFÍA 78
ANEXOS 81

TABLA DE ILUSTRACIONES

Ilustración 1. Algoritmo Teorema de Bayes 20
Ilustración 2. Funcionamiento Random Forest.....22
Ilustración 3. Aplicación máquinas de soporte vectorial.....23
Ilustración 4. Ecuación regresión lineal simple. 24
Ilustración 5. Ecuación regresión lineal múltiple. 24
Ilustración 6: Modelo Regresión Logística 25
Ilustración 7. Función de probabilidad sigmoide 26
Ilustración 8. Función sigmoide.....26
Ilustración 9. Función de predicción regresión logistica multinomial.....28
Ilustración 10. Clasificador k-vecino más cercano. 29
Ilustración 11. Matriz de confusión.....30
Ilustración 12. Exactitud. Matriz de confusión. 31
Ilustración 13.Precisión: Matriz de confusión. 31
Ilustración 14. Sensibilidad: Matriz de confusión. 32
Ilustración 15. Especificidad: Matriz de confusión.....33
Ilustración 16. Ejemplo matriz de confusión multiclase.....34
Ilustración 17. Ecuaciones de precisión y Recall ejemplo.....34
Ilustración 18. Metodología aplicada predicción fútbol.....43
Ilustración 19. Obtención de información as.com Colombia. 47
Ilustración 20. Variables conjunto de datos.....49
Ilustración 21. Resultados fase construcción conjunto de datos.....49
Ilustración 22. Construcción variable dependiente.....50
Ilustración 23. Balance clases conjunto de datos. 51
Ilustración 24. Proporción de victorias equipo local, vs goles anotados. 52
Ilustración 25. Proporción de empates vs goles anotados.....53
Ilustración 26. Proporción de victorias equipo visitante vs goles. 53
Ilustración 27. Análisis posesión de balón. 54
Ilustración 28. Análisis tiros al arco.....55
Ilustración 29. Ecuaciones fuerza de ataque y defensa.....56
Ilustración 30. Análisis de correlación de datos. 57
Ilustración 31. Fuerza de ataque y defensa equipos del FPC.....59
Ilustración 32. Análisis de corrección de variables.....61
Ilustración 33. Método Hold-out. 62
Ilustración 34. Distribución del conjunto de datos. 63
Ilustración 35. Características conjunto entrenamiento datos 2019. 63
Ilustración 36. Ecuación de balanceo de clases. 65
Ilustración 37. Exactitud en función de la constante de penalización. 66
Ilustración 38. Matriz de confusión regresión logística multiclase. 67
Ilustración 39. Reporte de resultados regresión logística multiclase.....68
Ilustración 40. Precisión en función del parámetro n_estimator.....69
Ilustración 41. Matriz de confusión random forest.....69
Ilustración 42. Reporte de resultados random forest. 70
Ilustración 43. Exactitud en función de la constante de penalización. 71

Ilustración 44. Matriz de confusión máquinas de soporte vectorial. 72
Ilustración 45. Reporte de resultados máquinas de soporte vectorial. 73

TABLA DE TABLAS

Tabla 1. Tabla de características previas.....58
Tabla 2. Variables modelo de predicción. 60
Tabla 3. Tabla de resultados regresión logística multiclase. 67
Tabla 4. Tabla de resultados random forest. 70
Tabla 5. Tabla de resultados máquinas de soporte vectorial. 73
Tabla 6. Tabla general puntuación de precisión. 74
Tabla 7. Partidos por algoritmo. 75

LISTA DE ABREVIATURAS

ML: *Machine Learning.*

IA: Inteligencia artificial.

DIMAYOR: División Mayor del fútbol colombiano.

DIM: Deportivo Independiente Medellín.

FIFA: Federación internación de Fútbol Asociado.

RNN: *Recurrent neural network*.

RMSE: *Root Mean Squared Error*.

FACL: Fuerza de ataque como local.

FACV: Fuerza de ataque como visitante.

FDCL: Fuerza de defensa como local.

FDCV: Fuerza de defensa como visitante.

GMCL: Goles Marcados como local.

GMCV: Goles Marcados como visitante.

GRCL: Goles recibidos como local.

GRCV: Goles recibidos como visitante.

FTR: *Full Time Result* (Resultado tiempo completo).

PALABRAS CLAVE

Aprendizaje de máquina, fútbol, máquinas de soporte vectorial, *Random Forest*, Regresión logística.

GLOSARIO

ANÁLISIS DE DATOS: "*El proceso de obtener una comprensión de los datos mediante la consideración de muestras, mediciones y visualizaciones.*" … (developers.google.com, 2019)

CLASE: Valor de un conjunto de valores de segmentación enumerados para una etiqueta.

ENTRENAMIENTO: "*Proceso de determinar los parámetros ideales que conforman un modelo*." … (developers.google.com, 2019)

EXACTITUD: Fracción de predicciones que se realizaron correctamente en un modelo de clasificación.

HIPERPARAMETRO: Es un parámetro cuyo valor se establece antes de que comience el proceso de aprendizaje.

LOGIT: "*Vector de predicciones sin procesar que genera un modelo de clasificación, que comúnmente se pasa a una función de normalización."* … (developers.google.com, 2019)

PESO: Coeficiente para un atributo en un modelo de predicción.

PRECISIÓN: **"***Métrica para los modelos de clasificación. La precisión identifica la frecuencia con la que un modelo predijo correctamente la clase positiva."* … (developers.google.com, 2019)

PREDICCIÓN: "*Resultado de un modelo cuando se le proporciona un ejemplo de entrada*." … (developers.google.com, 2019)

SOBRE-MUESTREO: Agregar más muestras a clases subrepresentadas.

SUBMUESTREO: Eliminar muestras de clases sobre representadas.

VALORES ATÍPICOS: Valores distantes de la mayoría de los demás valores

RESUMEN

En el presente trabajo de investigación tecnológica utilizó algoritmos de aprendizaje de máquina para predecir quién será el ganador un partido de la categoría A del fútbol profesional colombiano, para ello, se realizó la obtención y construcción del conjunto de datos, dichos datos fueron obtenidos de la página Diario AS[1]. Posteriormente, se realizó un procesamiento, limpieza y transformación a la información que se obtuvo, luego, se realizó un análisis estadístico a las variables para identificar aquellas que serán tenidas en cuenta para ser utilizadas por los algoritmos de aprendizaje de máquina, después, se realizó muestreo de información siguiendo el método *Hold-Out* el cual consiste en dividir el conjunto de datos, uno de entrenamiento y otro de prueba, siguiendo la proporción 75% - 25% respectivamente.

Una vez realizado lo anterior, se procede a implementar los algoritmos de aprendizaje máquina *random forest*, Máquinas de soporte vectorial y regresión logística multiclase, los cuales generará una serie de clasificaciones marcadas por las clases victoria para el equipo visitante, victoria para el equipo local o un empate. Después, se realizó la medición del desempeño de cada uno de los algoritmos de aprendizaje de máquina a través de la métrica matriz de confusión, al final se discuten los resultados obtenidos por cada algoritmo, donde se identificó que el algoritmo que tuvo un mejor desempeño fue *random forest*.

[1] *Diario deportivo Diario AS* Obtenido de: https://colombia.as.com/

INTRODUCCIÓN

Con el avance de las tecnologías de la información y la comunicación, el análisis de datos (o minería de datos) se ha convertido en una parte fundamental en los negocios, en sectores como el comercio, la banca, ciencias sociales, medicina, astronomía (ICEMD, 2017). La minería de datos ahora está siendo utilizada en los deportes, en especial en el deporte más popular del mundo: El fútbol. Las opciones de aplicación de la minería de datos y los algoritmos de aprendizaje de máquina en el fútbol son diversas ya que comprenden el sector de la industria de las apuestas deportivas; el análisis del desempeño de los jugadores que conforman el equipo; la formación táctica del equipo y plan de juego; la cantidad de goles anotados por encuentro hasta el posible ganador antes de que comience el encuentro. (Ganesan, 2018).

Actualmente en Colombia, especialmente en el fútbol profesional colombiano no se hace uso de algoritmos de aprendizaje de máquina en ninguna de las competiciones que posee la División Mayor del fútbol colombiano (DIMAYOR), un sector donde el aprendizaje de máquina puede ser de gran utilidad para los equipos de fútbol colombiano, para así, generar ventajas competitivas sobre sus rivales de campo, es por esta razón, que este trabajo pretende utilizar los datos históricos de los partidos de fútbol de la categoría A del fútbol profesional colombiano, comprendido entre los años 2015 a 2018 dado que es el formato actual que tiene esta competición[1], para poder procesar esta información y así predecir quién será el posible ganador de un partido de fútbol.

Se utilizará la metodología de aprendizaje de máquina comprendido por los pasos de construcción del conjunto de datos, categorización de los datos, posteriormente un preprocesamiento de la información y limpieza de los datos, luego se realizará la extracción de características acompañadas por muestreo y validación de datos que permitan la construcción del modelo de predicción que finalmente tendrá como salida una categorización en las clases gana el equipo local (L), gana el equipo visitante (V) o hay un empate (E). El resultado de este experimento basado en aprendizaje de máquina tiene como fin apoyar la toma de decisiones de los equipos del fútbol profesional colombiano para generar un fútbol más competitivo, por otra parte, puede ser de utilidad para las empresas patrocinadoras para ver la eficiencia partido tras partido de los equipos de fútbol para promover sus productos y apoyar de forma económica a los equipos del fútbol profesional colombiano. (Marin, 2017)

[1] *Los cambios del Fútbol Profesional Colombiano en la temporada 2015*: https://www.winsports.co/futbol-colombiano/liga-aguila/noticias/los-cambios-del-futbol-profesional-colombiano-en-la-temporada-2015-34295

0. FICHA TÉCNICA

Fecha: 2019-05-30.
Encabezado: Proyecto de grado.
Título Modelamiento predictivo del ganador de un partido de fútbol de la categoría A del fútbol profesional colombiano usando aprendizaje de máquina.
Alternativa: Trabajo de investigación tecnológica.
Línea de investigación: Aprendizaje de máquina.

1. GENERALIDADES

1.1. ANTECEDENTES

Betegy es una empresa tecnología de predicción de fútbol de origen polaco, que ofrece servicios en línea de análisis de datos, "*proporciona predicciones precisas para más de 29 ligas y copas (más de 8 000 juegos durante el año). Analiza 50,000 puntos de datos de fútbol y los combina con modelos matemáticos y scripts de aprendizaje automático para producir las predicciones más precisas disponibles.*" (betegy, 2012) El algoritmo de *Betegy* se basa en diferentes fuentes de información como noticias públicas, resultados de partidos anteriores, las posiciones actuales de la liga, los goles anotados, modelos estadísticos y funcionalidades de redes neurales en consideración a muchas otras fuentes.

Opta es una compañía internacional fundada en el reino unido en el año 1996, es considerado como un proveedor mundial de datos deportivos detallados, suministrados a clientes de diferentes industrias como a medios de comunicación, casas de apuestas deportivas, equipos profesionales de fútbol, y empresas dedicadas al patrocinio y activación de las marcas. Desde su fundación en el año 1996 *Opta* cuenta con varias oficinas a nivel internacional, lo que garantiza el poder contar con expertos locales en materia deportiva.

Dentro de los diversos servicios ofrecidos por *Opta*, se encuentran la analítica predictiva, la cual cuenta con robustos modelos predictivos impulsados por datos históricos detallados utilizando métricas predictivas con fines de previsualización de torneos, participación con clientes de apuestas deportivas y la generación de debates sociales en medios de comunicación. (optasports, 2018)

Stats es una compañía estadounidense fundada por John Dewan en el año 1981 con sede Chicago, Illinois, ofrece servicios de datos deportivos y analítica de datos orientados a ligas y equipos deportivos para mejorar el rendimiento del equipo a través del seguimiento a jugadores, monitoreo y desarrollo de atletas y soluciones de análisis de vídeo mediante el uso de aprendizaje automático de máquina y la inteligencia artificial. (stats.com, 2018)

1.2. PLANTEAMIENTO DEL PROBLEMA

1.2.1 Descripción del problema. La historia del deporte más popular del planeta como se conoce hoy en día tiene sus orígenes hacia el año de 1863, cuando en Inglaterra se separaron los caminos del "Rugby-Football", permitiendo así la fundación de la asociación más antigua del mundo: la "Football Association" (Asociación de fútbol de Inglaterra), el primer órgano gubernamental del deporte (FIFA, 2018). Desde entonces el fútbol ha tenido un crecimiento constante, hasta ser el deporte más popular del mundo con un aproximado de 270 millones personas involucradas, según estadísticas del año 2006 (FIFA, 2006).

Con el pasar del tiempo, el fútbol se ha convertido en un deporte realmente competitivo y táctico, donde cada variable puede cambiar el curso del juego (Mehta, 2017). Los clubes de fútbol y los directivos deportivos siempre han pretendido buscar una ventaja competitiva frente a sus competidores, es por eso que los orígenes de la predicción de partidos de fútbol se realizaban principalmente por intuición o sensación de "intuición" (Ekefre, 2016). Poco después, se implementó el primer modelo "científico" e investigativo para predecir los resultados de los partidos de fútbol, el *modelo de Poisson* (Maher, 1982). Desde la década de los 60', el término "*aprendizaje automático*" toma fuerza en el área de la informática, más adelante en la década de los 90' el aprendizaje automático cambia su enfoque del conocimiento a un enfoque basado en los datos. "*Los científicos comienzan a crear programas para computadoras para analizar grandes cantidades de datos y sacar conclusiones, o "aprender", a partir de los resultados.*" … (Marr, 2016).

En la actualidad, los algoritmos de aprendizaje de máquina y la minería de datos tienen una participación importante en deportes como el baloncesto, fútbol americano y especialmente en el fútbol, con aplicaciones en el rendimiento físico de los jugadores implementadas por compañías tecnológicas como OPTA o STATS los cuales recopilan datos como actividades de mapa de calor, datos históricos, y el rendimiento de los jugadores dotándolos con acelerómetros, sensores de ritmo cardíaco, sistemas GPS para definir la mejor estrategia de juego, comprar o vender jugadores o evitar posibles lesiones en los jugadores (Pérez, 2017).

Actualmente en Colombia, el uso de algoritmos y metodologías de aprendizaje de máquina aplicado al deporte, en especial al fútbol es escaso, lo cual genera opiniones subjetivas por parte de la mayoría de los hinchas y la falta de profundidad por parte de la prensa y periodistas deportivos, dada la situación anteriormente expuesta, se puede ver en la semifinal de vuelta entre el equipo Deportes Tolima enfrentando al Deportivo Independiente Medellín (DIM), donde la mayoría de hinchas de cuadro "*piajo*" daban como ganador a su equipo por ser el campeón del

torneo apertura 2018 y por jugar en condición de local, aquella noche del 25 de noviembre de 2018 el Deportivo Independiente Medellín derroto en condición de visitante al cuadro Deportes Tolima 2 goles a 0, acabando con el favoritismo del cuadro "piajo", como lo expresa el columnista Amado Hernández Gaviria de la Liga Deportiva Postobón: "*Los favoritismo solo son teorías sustentadas en la subjetividad.*" (Gaviria, 2018) En aquella nota el centrocampista del equipo antioqueño Andrés Ricaurte menciona: "*Esto no es de favoritos. Pienso que el fútbol colombiano es muy parejo y puede suceder cualquier cosa*" [1].

En esta situación se enmarca un problema, dado que el fútbol es un deporte regido por emociones y sentimientos, esto genera contenido opinativo con tendencia a la subjetividad por parte de la prensa y periodistas deportivos y no agrega valor de análisis e investigación que a los hinchas del fútbol colombiano verdaderamente les interesa, esto podría cambiar si se implementará análisis a los datos que genera el fútbol colombiano y determinar con objetividad y claridad quién realmente seria el posible ganador de un partido.

Por otra parte, los equipos del fútbol profesional colombiano "*trabajan por alcanzar reconocimiento y nombre en cuanto a historial, prestigio y credibilidad*" (Marin, 2017) para lograr esto es importante sumar ingresos y patrocinios por parte de empresas, pero es en este punto donde los equipos del fútbol colombiano tienen dificultades ya que para las empresas es importante saber con anterioridad el rendimiento de un equipo de fútbol. En la medida en que un club sea más exitoso en resultados, tendrá más ganancias económicas y su patrimonio se fortalecerá[2] y se verá reflejado en la adquisición de más y mejores jugadores para así crear una liga de fútbol competitiva y de nivel.

1.2.2 Formulación del problema. Dada la descripción del problema en el ítem anterior se plantea la siguiente pregunta:

¿Pueden los algoritmos de aprendizaje de máquina predecir el ganador de un partido de la categoría A del fútbol profesional colombiano, basado en la información de los resultados de los años 2015 a 2018?

[1] *En el fútbol colombiano puede suceder cualquier cosa*. Disponible en: https://ligadeportiva.com/andres-ricaurte-puede-suceder-cualquier-cosa/

[2] *Así se mueven las marcas en el fútbol colombiano*. Disponible en: http://www.elcolombiano.com/deportes/futbol/las-marcas-de-futbol-mas-valiosas-de-colombia-AM6817727

1.3. OBJETIVOS

1.3.1 Objetivo General. Implementar un modelo de aprendizaje de máquina para predecir el ganador de un partido de la categoría A del fútbol profesional colombiano con base en la información de los resultados de los años 2015 a 2018.

1.3.2 Objetivos específicos.

- Identificar las variables para la construcción del modelo de predicción del ganador de un partido de la categoría A del fútbol profesional colombiano.

- Diseñar una estrategia metodológica basado en aprendizaje de máquina para la predicción del ganador de un partido de fútbol de la categoría A del fútbol profesional colombiano.

- Implementar un algoritmo basado en aprendizaje de máquina para la predicción del ganador de un partido del fútbol profesional colombiano.

- Medir el desempeño del modelo de predicción en base a la métrica de evaluación matriz de confusión para determinar la precisión del ganador de un partido de la categoría A del fútbol profesional colombiano.

1.4. JUSTIFICACIÓN

El aprendizaje de máquina es una de las metodologías inteligentes que han mostrado resultados prometedores en los dominios de clasificación y predicción. Una de las áreas de expansión que requiere una buena precisión predictiva es la predicción deportiva, debido a las grandes cantidades monetarias involucradas en las apuestas deportivas (Bunker & Thabtah, 2017). Además, los gerentes y dueños de los clubes se esfuerzan por obtener modelos de clasificación que puedan entender y así, formular estrategias necesarias para poder ganar partidos.

Actualmente en Colombia, el fútbol profesional colombiano no cuenta con aplicaciones de aprendizaje de máquina que genere valor agregado a la Liga Águila, a los equipos profesionales de fútbol, a sus diversos hinchas y a las empresas que ven con buenos ojos a los equipos para patrocinar sus productos y activar las marcas. La aplicación de un proyecto de aprendizaje de máquina enfocado al fútbol profesional colombiano significaría el crecimiento a nivel táctico de un equipo de fútbol, ya que permitirá reevaluar las tácticas implementadas para obtener mejores resultados, por otra parte en materia económica, para las empresas y patrocinadores seria de importancia conocer de forma anticipada el resultado de un partido de fútbol y realizar distintos escenarios para ver si un equipo es rentable para patrocinar su marca o producto y así, de esta forma permita crear una simbiosis que permita tanto el crecimiento de la empresa y el equipo de fútbol y genere una liga de fútbol más competitiva y atractiva a nivel internacional[1].

[1] *Liga Águila, la cuarta mejor del mundo FÚTBOL Según el ranking de la IFFHS.* Obtenido de: http://co.marca.com/claro/futbol/2019/01/28/5c4f619f268e3ed0208b45ab.html

1.5. ALCANCES Y LIMITACIONES

1.5.1 Alcance. Se implementará un modelo de aprendizaje de máquina para predecir el ganador de un partido de la categoría A del fútbol profesional colombiano, cuyo resultado será categorizado en tres clases: (L) victoria equipo local, (E) empate y (V) victoria equipo visitante.

1.5.2 Limitaciones. Debido a los cambios que tuvo la categoría A del fútbol profesional colombiano de cara al año 2015, se optó por acotar el conjunto de datos a partir de dicho año tanto torneo de apertura como finalización, en cuanto a la selección de la categoría A del fútbol profesional colombiano, se tuvo en cuenta la nómina limitada de jugadores con la que cuentan los equipos que conforman la liga principal de la división mayor del fútbol colombiano.

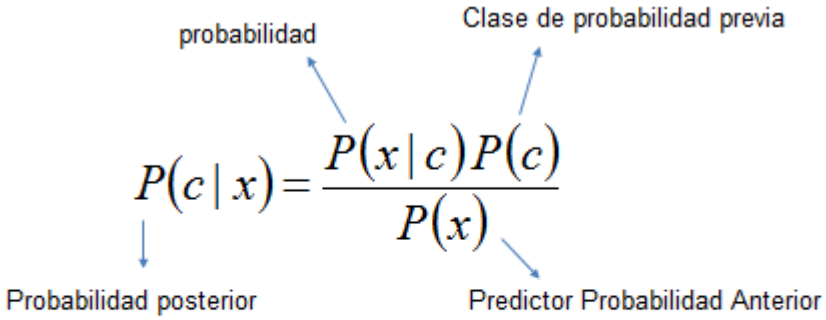
Para la predicción de partidos de fútbol de la categoría A del fútbol profesional colombiano del año 2019, no se tomaron en cuenta aquellos registros en los cuales figura el equipo Unión Magdalena, ya que este equipo, descendió desde el año 2005, por ende, no existen datos de entrenamiento para dicho equipo de fútbol ya que no se recolectan datos del año 2005 y tampoco se recolectan datos de la segunda división del fútbol profesional colombiano.

1.6. MARCO REFERENCIAL

1.6.1 Marco teórico.

1.6.1.1 Algoritmo Naïve Bayes. Es una técnica de clasificación basada en el teorema de Bayes con un supuesto de independencia entre los predictores. En términos simples, un clasificador Naive Bayes asume que la presencia de una característica particular en una clase no está relacionada con la presencia de cualquier otra característica. Junto con la simplicidad, se sabe que Naive Bayes supera incluso a los métodos de clasificación altamente sofisticados. (Soni, 2017)

Ilustración 1. Algoritmo Teorema de Bayes



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fuente: analytics vidhya, (2017). 6 *Easy Steps to Learn Naive Bayes Algorithm* [Ilustración]. Recuperado de https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Con base en la ilustración 1, a continuación, se describe la ecuación:

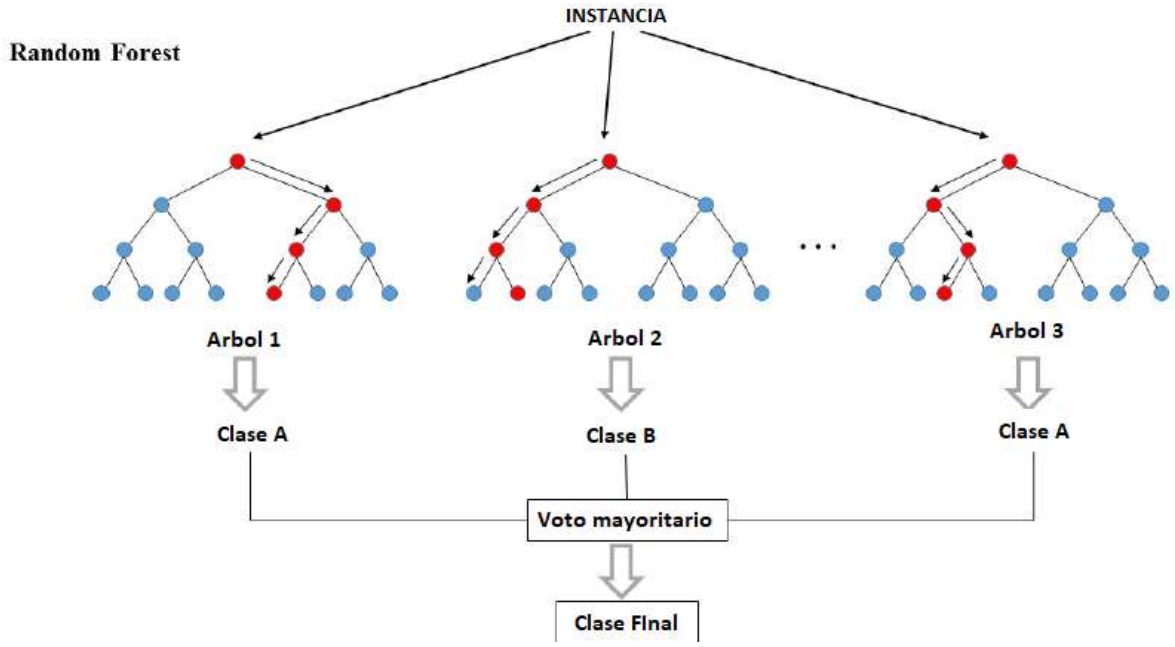
- P (c | x) es la probabilidad posterior de la clase (c, objetivo) dado predictor (x, atributos).
- P (c) es la probabilidad previa de clase.
- P (x | c) es la probabilidad que es la probabilidad de predictor de una clase determinada.
- P (x) es la probabilidad previa de predictor.

1.6.1.2 Algoritmo Random Forest. *"El algoritmo Random Forest fue desarrollado por el difunto profesor de Berkeley, Leo Breiman y Adele Cutler. Los bosques aleatorios generan su secuencia de modelos entrenándolos en subconjuntos de los datos. Los subconjuntos se extraen al azar del conjunto de entrenamiento completo. Una forma en la que se selecciona el subconjunto es muestrear aleatoriamente filas con reemplazo de la misma manera que el algoritmo de agregación bootstrap de Brieman. El otro elemento aleatorio es que los conjuntos de entrenamiento para los árboles individuales en el conjunto de bosques aleatorios no incorporan todos los atributos, sino que también toman un subconjunto aleatorio de los atributos."* (Bowles, 2015)

Para entender mejor este algoritmo, es necesario entender el componente fundamental sobre el cual se centra este algoritmo, los arboles de decisión. Un árbol de decisión es un algoritmo de aprendizaje automático capaz de ajustar conjuntos de datos complejos y realizar tareas de clasificación y regresión. La idea detrás de un árbol es buscar un par de valores variables dentro del conjunto de entrenamiento y dividirlo de tal manera que genere los mejores dos subconjuntos secundarios. El objetivo es crear ramas y hojas basadas en un criterio de división óptimo, un proceso llamado crecimiento de árboles. Específicamente, en cada rama o nodo, una declaración condicional clasifica el punto de datos basado en un umbral fijo en una variable específica, dividiendo así los datos. Para hacer predicciones, cada nueva instancia comienza en el nodo raíz y se mueve a lo largo de las ramas hasta que llega a un nodo de hoja donde no es posible realizar más ramificaciones. (medium.com, 2017)

Random Forest crea un conjunto de "*árboles de decisión y los combina para obtener una predicción más precisa y estable, agrega aleatoriedad adicional al modelo, mientras crece los árboles. En lugar de buscar la característica más importante al dividir un nodo, busca la mejor característica entre un subconjunto aleatorio de características. Esto da como resultado una amplia diversidad que generalmente resulta en un mejor modelo*" … (Gónzalez, 2019). En la ilustración 2 se puede ver un ejemplo de cómo funciona:

Ilustración 2. Funcionamiento Random Forest.

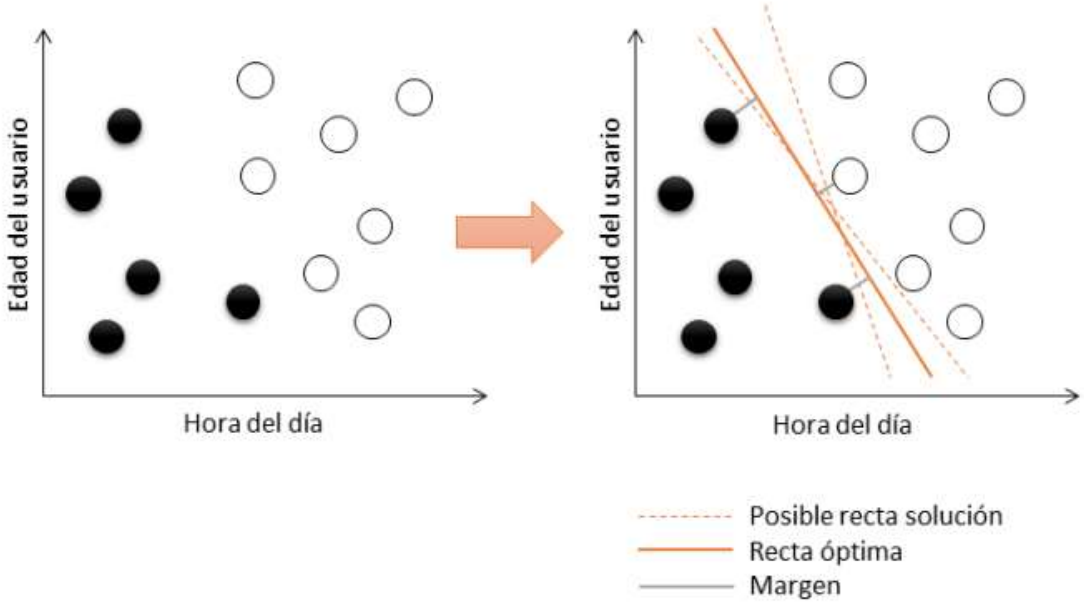


Fuente: medium.com, (2017). *Random Forest Simple Explanation* [Ilustración]. Recuperado de https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

1.6.1.3 Máquinas de Soporte Vectorial. Las máquinas de vectores de soporte (SVM) son un conjunto de métodos relacionados para el aprendizaje de máquina supervisado, aplicables tanto a problemas de clasificación como de regresión. Desde la introducción del clasificador SVM hace una década ganó popularidad debido a su sólida base teórica. El aprendizaje de máquina de vectores de soporte fue desarrollado por Vapnik et al. (Scholkopf et al., 1995, Scholkopf 1997) para implementar constructivamente los principios de la teoría del aprendizaje estadístico. En el marco del aprendizaje estadístico, el aprendizaje significa estimar una función a partir de un conjunto de ejemplos (los conjuntos de entrenamiento). Para hacer esto, una máquina de aprendizaje debe elegir una función de un conjunto dado de funciones, lo que minimiza un cierto riesgo (el riesgo empírico) de que la función estimada sea diferente de la función real (aún desconocida). El riesgo depende de la complejidad del conjunto de funciones elegidas, así como del conjunto de entrenamiento. Por lo tanto, una máquina de aprendizaje debe encontrar el mejor conjunto de funciones, según lo determinado por su complejidad, y la mejor función en ese conjunto.

"*Las SVM se empezaron a emplear para resolver problemas de clasificación y reconocimiento de patrones para luego extenderse en el estudio de predicción de series de tiempo. Los problemas de clasificación se emplean para obtener resultados de tipo cualitativo, por ejemplo, determinar la clase de un dato de entrada o características, mientras que las de tipo regresión son más útiles en problemas cuantitativos, cuando se trata de obtener una salida numérica al dato de entrada.*" … (Anzola, 2015) En la ilustración 3 se puede gráficamente el funcionamiento y aplicación de las máquinas de soporte vectorial.

Ilustración 3. Aplicación máquinas de soporte vectorial



Fuente: analiticaweb, (2016). Machine Learning y Support Vector Machines [Ilustración]. Recuperado de https://www.analiticaweb.es/machine-learning-y-support-vector-machines-porque-el-tiempo-es-dinero-2/

1.6.1.4 Regresión lineal. La regresión lineal es un tipo de análisis predictivo básico y de uso común. La regresión lineal intenta modelar la relación entre dos variables ajustando una ecuación lineal a los datos observados. Una variable se considera una variable explicativa y la otra se considera una variable dependiente. La representación de la regresión lineal es una ecuación que describe una línea que mejor se ajusta a la relación entre las variables de entrada (x) y las variables de salida (y), al encontrar ponderaciones específicas para las variables de entrada llamadas coeficientes (B).

El objetivo de la regresión lineal es aprender un modelo lineal en el que se pueda predecir (Y), mientras se intenta reducir el error. Al reducir el error, se aumenta inversamente la precisión del modelo de predicción. De ese modo, mejorando la función.

- Regresión lineal simple. Sólo se maneja una variable independiente, por lo que sólo cuenta con dos parámetros. Son de la forma como se observa en la ilustración 4:

Ilustración 4. Ecuación regresión lineal simple.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Fuente: (Arias, 2019)

- Regresión lineal múltiple. La regresión lineal permite trabajar con una variable a nivel de intervalo o razón. De la misma manera, es posible analizar la relación entre dos o más variables a través de ecuaciones, lo que se denomina regresión múltiple o regresión lineal múltiple.

  Constantemente en la práctica de la investigación estadística, se encuentran variables que de alguna manera están relacionadas entre sí, por lo que es posible que una de las variables pueda relacionarse matemáticamente en función de otra u otras variables. Se expresa de la siguiente forma:

Ilustración 5. Ecuación regresión lineal múltiple.

$$Y_i = \beta_0 + \sum \beta_p X_{pi} + \varepsilon_i$$

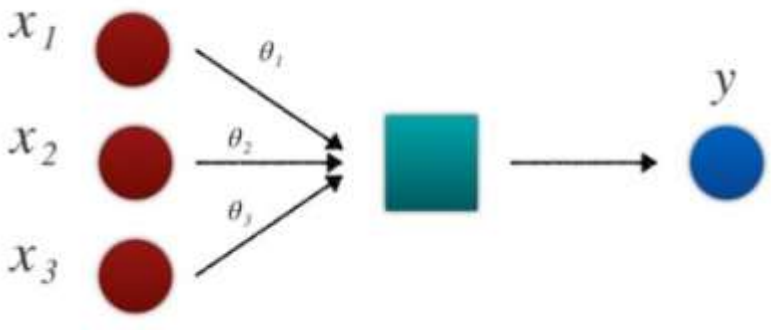
Fuente: (Arias, 2019)

1.6.1.5 Regresión Logística Binaria. La regresión logística es un método estadístico para analizar un conjunto de datos en el que hay una o más variables independientes que determinan un resultado. El resultado se mide con una variable dicotómica (en la que solo hay dos resultados posibles). Se utiliza para predecir un resultado binario dado un conjunto de variables independientes. Para representar el resultado binario o categórico, se utilizan variables ficticias. También la regresión logística se puede expresar como un caso especial de regresión lineal, es decir, cuando la variable de resultado es categórica, donde se está utilizando el registro de probabilidades como variable dependiente. En palabras simples, predice la probabilidad de ocurrencia de un evento al ajustar los datos a una función logit1. (ml-cheatsheet, 2017)

La regresión logística fue desarrollada por el estadístico David Cox en 1958. Este modelo logístico binario se usa para estimar la probabilidad de una respuesta binaria basada en una o más variables (características) predictoras (o independientes). Permite decir que la presencia de un factor de riesgo aumenta la probabilidad de un resultado dado en un porcentaje específico.

Como todos los análisis de regresión, la regresión logística es un análisis predictivo. La regresión logística se utiliza para describir datos y para explicar la relación entre una variable binaria dependiente y una o más variables independientes nominales, ordinales, de intervalo o de relación. (Gandhi, towardsdatascience.com, 2018) En la ilustración 6 se observa un modelo de regresión logística.

Ilustración 6: Modelo Regresión Logística



Fuente: towardsdatascience, (2017). Machine Learning Part 3: Logistic Regression [Ilustración]. Recuperado de https://towardsdatascience.com/machine-learning-part-3-logistics-regression-9d890928680f

[1] *¿Qué es una función Logit y por qué usar la regresión logística?:* https://www.theanalysisfactor.com/what-is-logit-function/

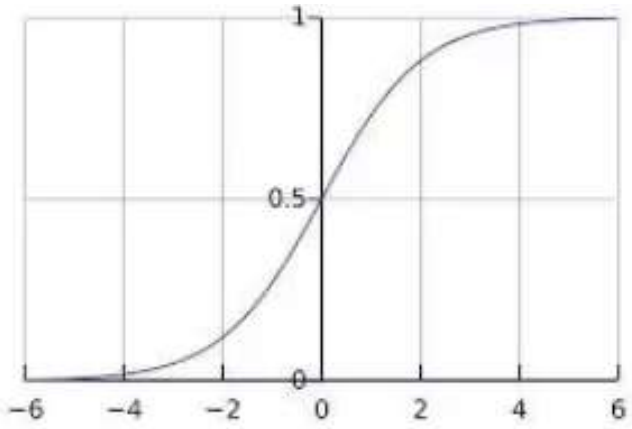
Esta función de probabilidad es la ' Función sigmoidal' como se observa en la ilustración 7:

Ilustración 7. Función de probabilidad sigmoide

$$\frac{1}{1 + e^{(-z)}}$$

En la ilustración 8 se muestra gráficamente la función, se vería de la siguiente forma:

Ilustración 8. Función sigmoide.



Fuente: quora.com, (2014). What exactly is a logistic regression algorithm in machine learning. What are its applications? [Ilustración]. Recuperado de https://www.quora.com/What-exactly-is-a-logistic-regression-algorithm-in-machine-learning-What-are-its-applications

1.6.1.6 Regresión Logística Multinomial. La regresión logística multinomial es un método de clasificación que generaliza la regresión logística a problemas a problemas multiclase, es decir, con más de dos posibles resultados discretos. Es un modelo que se utiliza para predecir las probabilidades de los diferentes resultados posibles de una variable dependiente distribuida categóricamente, dado un conjunto de variables independientes. La regresión logística multinomial es una solución particular para los problemas de clasificación que utilizan una combinación lineal de las características observadas y algunos parámetros específicos del problema para estimar la probabilidad de cada valor particular de la variable dependiente[1].

Al utilizar la regresión logística multinomial es apropiado verificar una serie de suposiciones sobre los datos, los cuales son:

- Supuesto 1: La variable dependiente debe medirse de manera nominal, es decir, Una variable puede ser tratada como nominal cuando sus valores representan categorías que no obedecen a una clasificación intrínseca[2].

- Supuesto 2: La variable independiente son continuas, ordinales o nominales. Sin embargo, es decir, una variable puede ser tratada como ordinal cuando sus valores representan categorías con alguna clasificación intrínseca. Las variables independientes ordinales deben tratarse como continuas o categóricas.

- Supuesto 3: Debe tener independencia de observaciones y la variable dependiente debe tener categorías mutuamente excluyentes y exhaustivas.

- Supuesto 4: No debe haber multicolinealidad. La multicolinealidad ocurre cuando se tiene dos o más variables independientes que están altamente correlacionadas entre sí.

- Supuesto 5: Debe haber una relación lineal entre las variables independientes continuas y la transformación *logit* de la variable dependiente.

- Supuesto 6: No debe haber valores atípicos, valores de alto apalancamiento o puntos altamente influyentes[3].

---

[1] *Regresión logística multinomia*l Obtenido de: https://en.wikipedia.org/wiki/Multinomial_logistic_regression

[2] *Nivel de medición de variables* Obtenido de: https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/base/dataedit_define_variable_measurement.html

[3] Assumptions Multinomial Logistic Regression Obtenido de: https://statistics.laerd.com/spss-tutorials/multinomial-logistic-regression-using-spss-statistics.php

Al igual que en otras formas de regresión lineal, la regresión logística multinomial utiliza una función predictiva lineal $f(k, i)$ para predecir la probabilidad de que la observación i tenga un resultado i, de la siguiente forma:

Ilustración 9. Función de predicción regresión logistica multinomial

$$f(k, i) = \boldsymbol{\beta}_k \cdot \mathbf{x}_i$$

Fuente: wikipedia.org, (2018). Regresión logística multinomial [Ilustración]. Recuperado de: https://en.wikipedia.org/wiki/Multinomial_logistic_regression

donde Xi es el vector de variables explicativas que describen la observación i, βk es un vector de ponderaciones correspondientes al resultado k, y la puntuación (Xi, k) es la puntuación asociada con la asignación de la observación i a la categoría k. En la teoría de la elección discreta, donde las observaciones representan personas y los resultados representan elecciones, la puntuación se considera la utilidad asociada con la persona i que elige el resultado k. El resultado predicho es el que tiene la puntuación más alta.

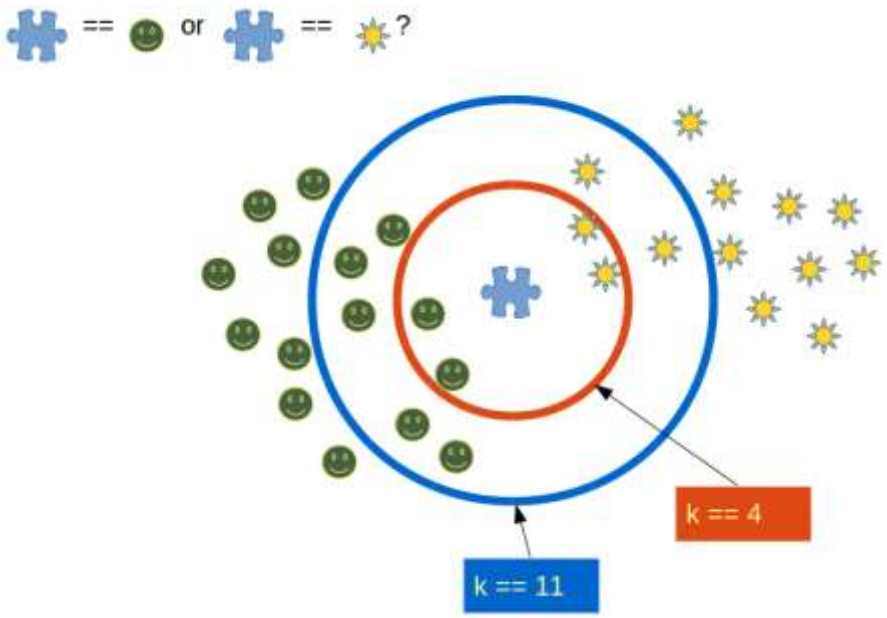
1.6.1.7 K nearest neighbor. Es uno de los algoritmos de clasificación más básicos pero esenciales en aprendizaje automático. Pertenece al dominio de aprendizaje supervisado y encuentra una aplicación intensa en el reconocimiento de patrones, la extracción de datos y la detección de intrusos. K-Nearest Neighbors (KNN) se puede utilizar para problemas de predicción tanto de clasificación como de regresión. Sin embargo, es más ampliamente utilizado en problemas de clasificación en la industria. Para evaluar cualquier técnica generalmente se fija 3 aspectos importantes:

- Facilidad para interpretar la salida.
- Tiempo de cálculo
- Poder predictivo

“K” significa algoritmo, para cada punto de datos de prueba, estaríamos viendo los K puntos de datos de entrenamiento más cercanos y tomaríamos las clases más frecuentes y asignaríamos esa clase a los datos de prueba. Por lo tanto, K representa el número de puntos de datos de entrenamiento que se encuentran cerca

del punto de datos de prueba que usaremos para encontrar la clase. (Gandhi, towardsdatascience.com, 2017). A continuación, se ilustra un ejemplo:

Ilustración 10. Clasificador k-vecino más cercano.

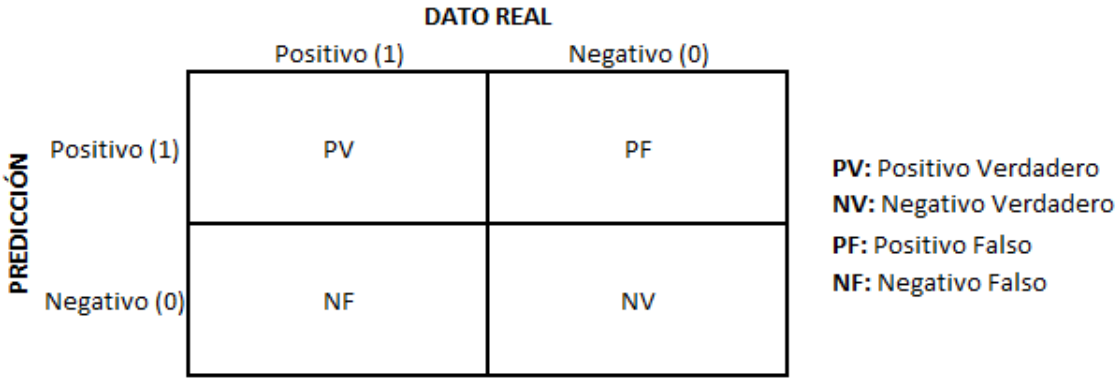


Fuente: python-course, (2018). Clasificador k-vecino más cercano [Ilustración]. Recuperado de https://www.python-course.eu/k_nearest_neighbor_classifier.php

1.6.1.8 Métricas de desempeño.

- Matriz de confusión. La matriz de confusión es una presentación práctica de la precisión de un modelo con dos o más clases. En la ilustración 11 se puede observar un ejemplo:

Ilustración 11. Matriz de confusión.



Fuente: (Arias, 2019)

Términos asociados con la matriz de confusión:

- Positivos verdaderos (PV). Los positivos verdaderos son un caso en donde el dato real es 1 (Verdadero) y la predicción también es 1 (Verdadero).

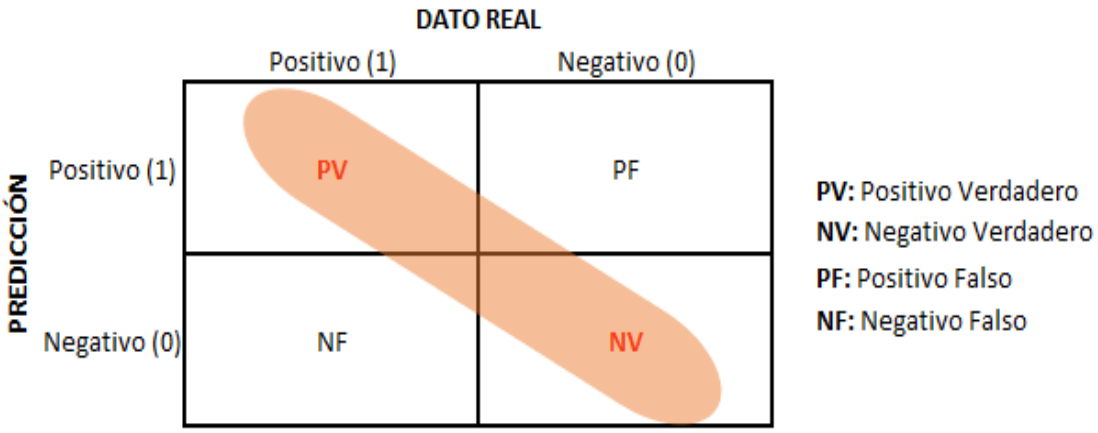
- Negativos verdaderos (NV). Los negativos verdaderos son un caso en donde el dato real es 0 (Falso) y el pronóstico también es 0 (Falso).

- Positivos Falsos (PF). Los positivos falsos son los casos en donde el dato real es 0 (Falso) y el pronóstico es 1 (Verdadero). Falso es porque el modelo ha pronosticado incorrectamente y positivo porque la clase predicha fue positiva. (1)

- Negativos falsos (NF). Los falsos negativos son los casos en donde el dato real es 1 (Verdadero) y el pronóstico es 0 (Falso). Falso es porque el modelo ha predicho incorrectamente y negativo porque la clase predijo que era negativa. (0).

- Exactitud. La precisión en los problemas de clasificación es el número de predicciones correctas realizadas por el modelo sobre todo tipo de predicciones realizadas. En la ilustración 12 se muestra un ejemplo con su respectiva formula:

Ilustración 12. Exactitud. Matriz de confusión.

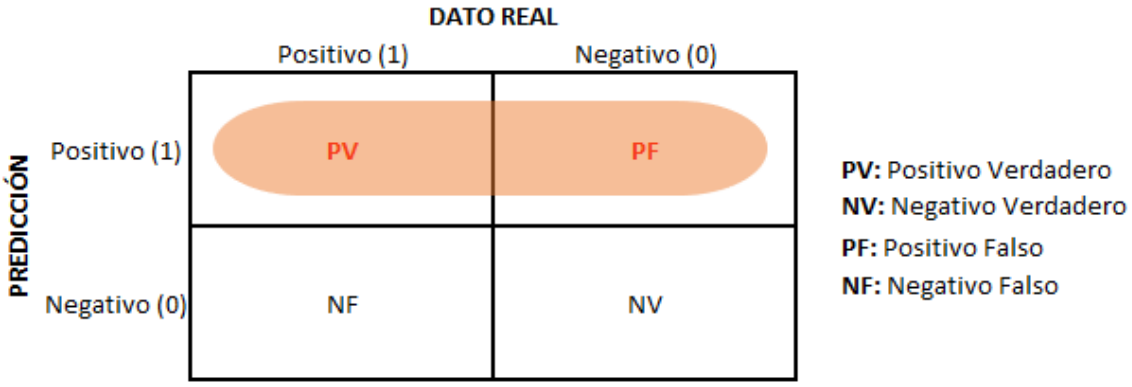


$$Exactitud = \frac{PV + NV}{PV + NV + NF + PF}$$

Fuente: (Arias, 2019)

- Precisión. La precisión es una medida que indica qué proporción de precisión tiene el modelo de predicción. Por ejemplo: qué proporción de pacientes a los que se diagnosticaron que tienen cáncer, en realidad tuvieron cáncer. En la ilustración 13 se observa un ejemplo con su respectiva formula:

Ilustración 13.Precisión: Matriz de confusión.

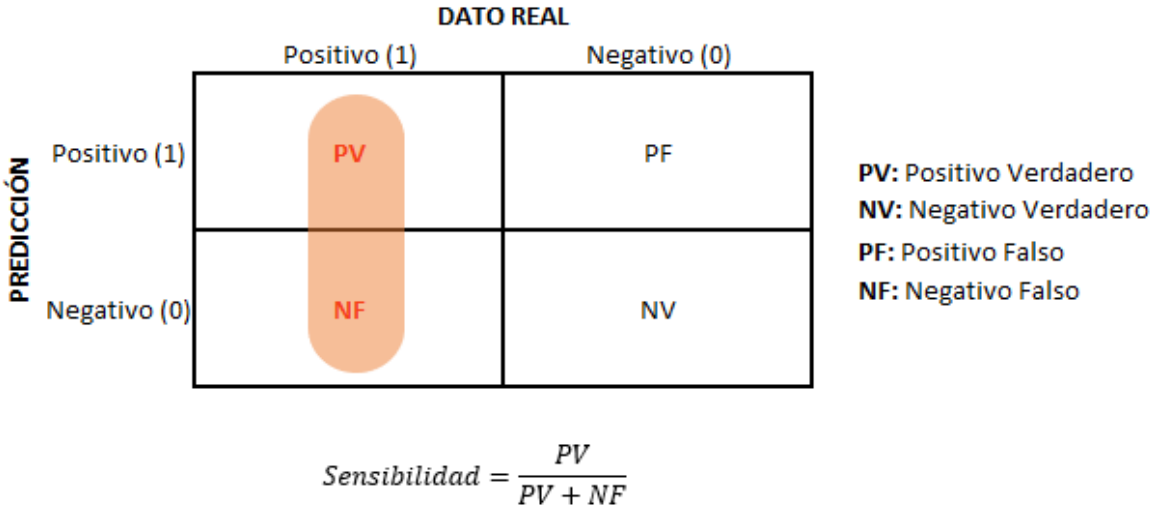


$$Precisión = \frac{PV}{PV + PF}$$

Fuente: (Arias, 2019)

- Recall o Sensibilidad. La proporción de casos positivos reales que están correctamente identificados. Por ejemplo: Qué proporción de pacientes que realmente tuvieron cáncer fue diagnosticado por el algoritmo como si tuviera cáncer. En la ilustración 14 se observa un ejemplo con su respectiva formula:

Ilustración 14. Sensibilidad: Matriz de confusión.

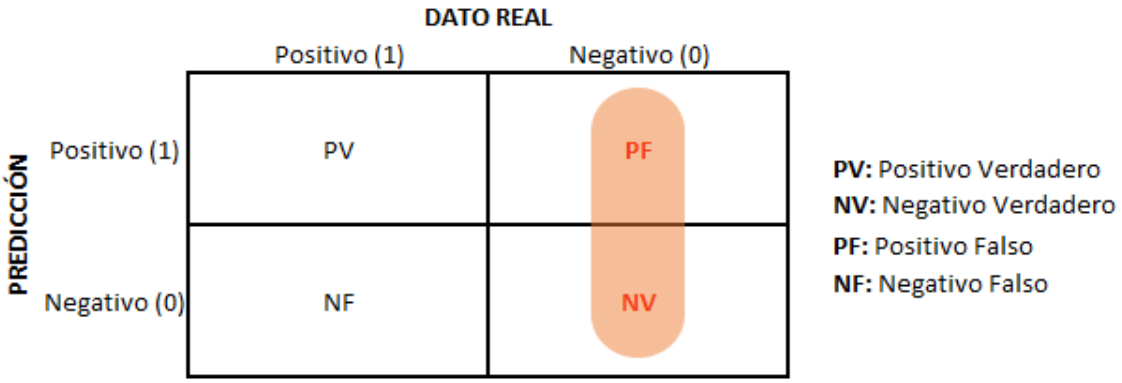


$$Sensibilidad = \frac{PV}{PV + NF}$$

Fuente: (Arias, 2019)

- Especificidad. La proporción de casos negativos reales que están correctamente identificados. Por ejemplo: qué proporción de pacientes que NO tuvieron cáncer, fueron predichos por el modelo como no cancerosos. En la ilustración 15 se observa un ejemplo con su respectiva formula:

Ilustración 15. Especificidad: Matriz de confusión.



$$Especificidad = \frac{NV}{PF + NV}$$

Fuente: (Arias, 2019)

En base a lo anterior, Para medir los resultados de los algoritmos de aprendizaje de máquina multiclase (más de dos categorías o clases predictoras), la matriz de confusión necesita una generalización para el caso multiclase.

- Precisión.

$$Precisión\ i = \frac{Mii}{\sum j\ Mji}$$

Fuente: (Arias, 2019)

la precisión es la fracción de casos en que el algoritmo predijo correctamente la clase i de todas las instancias en las que el algoritmo predijo i.

- Recall o sensibilidad.

$$Recall\ i = \frac{Mii}{\sum j\ Mij}$$

Fuente: (Arias, 2019)

Recall o también llamado sensibilidad, es la fracción de casos en los que el algoritmo predijo correctamente i de todos los casos que están etiquetados como i.

En la ilustración 16 y 17 se puede observar cada uno los conceptos anteriormente expuestos

Ilustración 16. Ejemplo matriz de confusión multiclase.

		Predicción		
		Avion	**Carro**	**Moto**
Reales	**Avion**	6	2	0
	Carro	1	6	0
	Moto	1	1	8

Fuente: (Arias, 2019)

Ilustración 17. Ecuaciones de precisión y *Recall* ejemplo.

$$Precisión\ avión = \frac{6}{6+1+1} = 0{,}75 \qquad recall\ avion = \frac{6}{6+2+0} = 0{,}75$$

$$Precisión\ carro = \frac{6}{2+6+1} = 0{,}67 \qquad recall\ avion = \frac{6}{1+6+0} = 0{,}86$$

$$Precisión\ moto = \frac{8}{0+0+8} = 1 \qquad recall\ moto = \frac{8}{1+1+8} = 0{,}8$$

Fuente: (Arias, 2019)

1.6.2 Marco Conceptual.

1.6.2.1 Fútbol. La historia del deporte más popular del planeta como se conoce hoy en día tiene sus orígenes hacia el año de 1848, cuando en Inglaterra se separaron los caminos del "Rugby-Football", permitiendo así la fundación de la asociación más antigua del mundo: la "Football Association" (Asociación de fútbol de Inglaterra), el primer órgano gubernamental del deporte (FIFA, 2018). Desde entonces el fútbol ha tenido un crecimiento constante, hasta ser el deporte más popular del mundo con un aproximado de 270 millones personas involucradas, según estadísticas del año 2006 (FIFA, 2006).

El fútbol es un deporte jugado entre dos equipos, cada uno con once jugadores, dentro del campo de juego también se encuentran 4 árbitros que se encargan de cumplir las normas del juego. El campo de juego es rectangular con una pelota esférica hasta que la pelota cruce un área de red llamada la meta. El objetivo del juego es anotar un gol dentro de la meta del equipo contrario. El equipo que marque más goles gana el juego. Si la puntuación entre ambos equipos está nivelada, se declara un empate, o se proporcionarán minutos adicionales para continuar jugando o hacer una tanda de penales para determinar el ganador, según el formato del juego. (sportscourtdimensions.com, 2015)

- División Mayor del fútbol colombiano (DIMAYOR). "*fue fundada el 26 de junio de 1948 y, de conformidad con su objeto estatutario, es la entidad que se encarga de organizar, administrar y reglamentar los campeonatos del Fútbol Profesional Colombiano.*" (dimayor.com.co, 2018)

- La Liga Águila. "*Certamen en el que compiten los 20 equipos de la categoría "A" y se coronan 2 campeones por año. Cada uno de ellos obtiene el título respectivo y adicionalmente un cupo en la Copa Libertadores de América del siguiente año.*" (dimayor.com.co, 2018)

- La Liga Profesional Femenina Águila. "*Es la competencia en donde participan los 23 clubes femeninos del FPC en un campeonato anual. El equipo campeón tiene el derecho a participar en la Copa Libertadores Femenina y es el representante de Colombia en la Copa DIMAYOR – LaLiga Women.*" (dimayor.com.co, 2018)

- El Torneo Águila. "*En el cual se enfrentan los 16 equipos de la categoría "B". Al final del año ascienden 2 clubes a la categoría superior y entran a competir en la Liga Águila de la siguiente temporada.*" (dimayor.com.co, 2018)

- La Superliga Águila. "*Enfrenta a los 2 campeones del año de la Liga Águila. En caso de que un mismo equipo gane los 2 eventos, la Superliga la disputa ese club, contra el mejor equipo de la tabla de reclasificación.* (dimayor.com.co, 2018)

1.6.2.2 Minería de datos. La minería de datos es el proceso de analizar patrones de datos ocultos desde diferentes perspectivas para la categorización en información útil, que se recopila y ensambla en áreas comunes como los almacenes de datos, para un análisis eficiente los algoritmos de minería de datos facilitan la toma de decisiones en los negocios y empresas los cuales son indispensables para reducir costos y aumentar los ingresos.[1]

La minería de datos también se conoce como descubrimiento de datos y descubrimiento de conocimiento.

Los principales pasos involucrados en un proceso de minería de datos son:

- Extraer, transformar y cargar datos en un almacén de datos.
- Almacenar y gestionar datos en una base de datos multidimensional.
- Proporcionar acceso a los datos a analistas de negocios que utilizan software de aplicación.
- Presentar los datos analizados en formas fácilmente comprensibles, como los gráficos.

1.6.2.3 Aprendizaje de máquina. El aprendizaje de máquina es una aplicación de inteligencia artificial (IA) que proporciona a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programado explícitamente. El aprendizaje de máquina se centra en el desarrollo de programas informáticos que pueden acceder a los datos, utilizarlos y aprender por sí mismos.

El proceso de aprendizaje comienza con observaciones o datos, como ejemplos de experiencia directa o instrucción, para buscar patrones en los datos y tomar mejores

[1] *Minería de datos*. Obtenido de: https://www.techopedia.com/definition/1181/data-mining

decisiones en el futuro en función de los ejemplos que se brinda al modelo. El objetivo principal del aprendizaje de máquina es permitir que las computadoras aprendan automáticamente sin intervención o asistencia humana y ajustar las acciones en consecuencia. [1]

- Aprendizaje supervisado. El aprendizaje supervisado es la tarea de minería de datos de "*inferir una función a partir de datos de entrenamiento etiquetados. Los datos de entrenamiento consisten en un conjunto de ejemplos de entrenamiento. En el aprendizaje supervisado, cada ejemplo es un par que consiste en un objeto de entrada (típicamente un vector) y un valor de salida deseado (también llamado señal de supervisión). Un algoritmo de aprendizaje supervisado analiza los datos de entrenamiento y produce una función inferida, que se puede utilizar para mapear nuevos ejemplos*" … (code.i-harness, 2016). Un escenario óptimo permitirá que el algoritmo determine correctamente las etiquetas de clase para instancias no vistas. Esto requiere que el algoritmo de aprendizaje generalice de los datos de entrenamiento a situaciones invisibles de una manera razonable.

- Aprendizaje no supervisado. "*El aprendizaje no supervisado estudia cómo los sistemas pueden inferir una función para describir una estructura oculta a partir de datos sin etiquetar. El sistema no encuentra la salida correcta, pero explora los datos y puede extraer inferencias de conjuntos de datos para describir estructuras ocultas de datos sin etiquetar*." (Varone, 2018)

- Aprendizaje semi-supervisado. Se ubican en algún lugar entre el aprendizaje supervisado y el no supervisado, ya que utilizan datos tanto etiquetados como no etiquetados para entrenamiento, generalmente una pequeña cantidad de datos etiquetados y una gran cantidad de datos no etiquetados. Los sistemas que utilizan este método pueden mejorar considerablemente la precisión del aprendizaje. Por lo general, el aprendizaje semi-supervisado se elige cuando los datos etiquetados adquiridos requieren recursos calificados y relevantes para entrenarlos.

- Aprendizaje automático de refuerzo. Son un método de aprendizaje que interactúa con su entorno produce acciones y descubre errores o recompensas. La búsqueda de prueba, error y recompensa diferida son las características más relevantes del aprendizaje por refuerzo. Este método permite que las máquinas y los agentes de software determinen

[1] *¿Qué es el aprendizaje automático?* Obtenido de: https://www.expertsystem.com/machine-learning-definition/

automáticamente el comportamiento ideal dentro de un contexto específico para maximizar su rendimiento. Se requiere un simple *feedback* de recompensa para que el agente sepa qué acción es mejor; Esto se conoce como la señal de refuerzo.

1.6.2.4 Aprendizaje profundo. El aprendizaje profundo es una función de inteligencia artificial que imita el funcionamiento del cerebro humano en el procesamiento de datos y la creación de patrones para su uso en la toma de decisiones. El aprendizaje profundo es un subconjunto del aprendizaje automático en Inteligencia Artificial (AI) que tiene redes capaces de aprender sin supervisión a partir de datos sin estructurar o sin etiquetar.

"*El aprendizaje profundo, un subconjunto del aprendizaje automático, utiliza un nivel jerárquico de redes neuronales artificiales para llevar a cabo el proceso de aprendizaje automático. Las redes neuronales artificiales se construyen como el cerebro humano, con nodos neuronales conectados entre sí como una red. Mientras que los programas tradicionales generan análisis con datos de forma lineal, la función jerárquica de los sistemas de aprendizaje profundo permite a las máquinas procesar datos con un enfoque no lineal.*" (Momoh, 2018)

1.6.3 Marco Legal.

1.6.3.1 Habeas Data. "*El Habeas Data es el derecho que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bancos de datos y en archivos de entidades públicas y privadas*". (dinero.com, 2013) Este derecho está regulado por el Artículo 15 de la Constitución Política de Colombia[1] y sancionada por el gobierno nacional a través de la ley 1266 de 2008[2].

La División Mayor del Fútbol Colombiano ha adoptado y acatado los lineamientos establecidos por dichas normas legales, ya que como entidad encargada de administrar y reglamentar los torneos de fútbol profesional en Colombia, es su deber principal velar por los derechos constitucionales que tienen todas las personas a conocer, actualizar y rectificar la información personal que haya sido recogida sobre ellas en bases de datos o archivos, correspondiente a clubes filiados, personal de

[1] *Articulo 15 Constitución política de Colombia* Obtenido de: http://www.constitucioncolombia.com/titulo-2/capitulo-1/articulo-15

[2] *LEY ESTATUTARIA 1266 DE 2008* Obtenido de: http://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=34488

los clubes afiliados, proveedores y trabajadores y cualquier persona perteneciente a dichos grupos de interés[1].

1.6.3.2 Aviso legal Diario AS. Parte del insumo utilizado en el presente trabajo de investigación tecnológica fue obtenido de la página web Diario AS[2], la cual proporcionó los datos y estadísticas de los partidos de la categoría A del fútbol profesional colombiano con fines netamente académicos. En la página del diario deportivo, se enuncian algunas de las políticas legales y de uso del portal[3], algunos de ellos como la propiedad intelectual e industrial que son importantes al momento de usar dicha información.

[1] *Política de privacidad y de protección de datos personales división mayor del fútbol colombiano ("DIMAYOR")* Obtenido de*:* http://dimayor.com.co/wp-content/uploads/2018/11/20171218-Poli%CC%81tica-de-privacidad-y-proteccio%CC%81n-de-datos-personales-2.pdf

[2] *Diario deportivo Diario AS* Obtenido de: https://colombia.as.com/

[3] *Aviso legal Diario AS* Obtenido de: https://as.com/diarioas/aviso_legal.html

1.7. ESTADO DEL ARTE

El proceso realizado en el estado del arte consistió en la lectura de diferentes publicaciones y revistas académicas entre los años 2014 a 2018, en el proceso se identificaron diferentes metodologías y algoritmos de aprendizaje de máquina para pronosticar en el fútbol, algunas de las investigaciones se muestran a continuación.

En la publicación realizada por los autores Engin Esme & Mustafa Servet Kiran para la revista *International Journal of Machine Learning and Computing*, Vol. 8, del año 2018, se propuso un modelo de predicción basado en el algoritmo *k-nearest neighbor* (algoritmo del vecino más cercano) para identificar el ganador de un partido de fútbol en la competición de la Superliga de Turquía utilizando un enfoque de diseño básico para medir la similitud entre las competencias basadas en las probabilidades de apuestas. Este modelo mejoró utilizando datos de juegos anteriores, disminuyendo significativamente el margen de error de los juegos predichos.

El conjunto de datos empleado para este estudio fue los resultados de las temporadas 2011 a 2016 de la Superliga turca, competencia en la cual participan 18 equipos de fútbol con una duración de 34 semanas. Los datos fueron separados de tal forma que una parte era un conjunto de datos entrenamiento y la otra mitad un conjunto de datos de prueba, se obtuvieron un total de 17 características, 10 de las cuales se basaron en las probabilidades de apuestas y 7 de ellas se basaron en otros datos estadísticos. Los datos para las competiciones se obtuvieron de www.mackolik.com[1]. El éxito de los resultados de predicción utilizando el modelo desarrollado se evaluó tomando como referencia las tasas de probabilidad del corredor de apuestas. Sin utilizar el análisis de riesgo, el modelo desarrollado hizo mejores predicciones a una tasa de 1.96% según las apuestas de resultados a tiempo completo y a una tasa de 7.84% según las apuestas de doble posibilidad. (Esme & Kiran, 2018)

El trabajo expuesto por el autor Anand Ganesan para la revista *International Journal of Pure and Applied Mathematics* Vol. 118 del año 2018, se empleó el uso de 3 algoritmos como *Support Vector Machines* (Máquinas de soporte vectorial), *XGBoost*[2] (eXtreme Gradient Boosting) y *Logistic Regression* (Regresión logística). Este modelo se aplica a los datos del equipo y los resultados de los encuentros recopilados en http://www.football-data.co.uk/ durante las últimas temporadas de la

[1] *Portal de resultados deportivos*: https://www.mackolik.com/
[2] *Boosting en Machine Learning*: https://relopezbriega.github.io/blog/2017/06/10/boosting-en-machine-learning-con-python/

Premier Legue. El modelo propuesto por el autor sugiere el entrenamiento de datos en los diferentes clasificadores de aprendizaje automático, posteriormente compara el rendimiento de cada clasificador y se selecciona el de mejor resultado, el resultado final del modelo propuesto por el autor será categorizado en tres clases, victoria del equipo local, victoria del equipo visitante o un empate. (Anand, 2018)

En el trabajo de investigación expuesto por el autor Abel Hijmans en el año 2016, se emplearon diversos algoritmos para predecir el ganador de un partido del fútbol holandés, entre los algoritmos utilizados se encuentran GBM (*Gradient Boosting Machine*) –modelo de árbol aleatorio-, *Naïve Bayes* y *k-nearest neighbor*. Se estudiaron los resultados de cada uno de los modelos expuestos y se analizaron las predicciones a profundidad concluyendo así en su investigación que el modelo GBM (modelo de árbol aleatorio) presentaba un mayor poder predictivo sobre los demás, con un 60.22% en promedio correctamente, después le sigue el modelo *Naïve Bayes* y el modelo *k-nearest neighbor* con una puntuación de predicción del 42% y el 58,62% respectivamente. (Hijmans, 2016)

Stratagem una compañía de tecnología financiera, proporciona servicios de investigación, análisis y comercio; y herramientas para el mercado del deporte. La compañía ofrece StrataBet, una plataforma de comercio deportivo que amplifica la inteligencia del comercio deportivo. La compañía también proporciona una base de datos de jugadores y equipos con calificación personalizada para mostrar el impacto de posibles eventos. Stratagem Technologies Limited se incorporó en 2012 y tiene su sede en Londres, Reino Unido.

"*Stratagem está utilizando redes neuronales profundas para lograr esta tarea, la misma tecnología que ha encantado a las empresas más grandes de Silicon Valley. Es un buen ajuste, ya que esta es una herramienta que es muy adecuada para analizar grandes cantidades de datos. Como señala Koukorinis, al analizar deportes, hay muchísimos datos para aprender. El software de la compañía actualmente está absorbiendo miles de horas de instalaciones deportivas para enseñarle patrones de fracaso y éxito, y el objetivo final es crear una IA que pueda ver una variedad de media docena de eventos deportivos diferentes simultáneamente en la televisión en vivo, extrayendo ideas. como lo hace*." (Vincent, 2017)

Numberfire es una plataforma de análisis de datos deportivos de origen estadounidense fundado por Nik Bonaddio en el año 2010, lleva los datos deportivos a nuevos niveles inteligentes. Al combinar métricas derivadas matemáticamente con algoritmos avanzados que tienen en cuenta las variables situacionales, *numberFire* convierte los "datos no estructurados y engañosos" en torno a los deportes en

estadísticas y predicciones muy precisas para los jugadores y equipos de la NFL, MLB y NBA. (numberfire.com, 2018)

NumberFire afirma que sus datos les dan a los usuarios un 31 por ciento más de posibilidades de ganar sus ligas de fantasía y supera las proyecciones proporcionadas por las ligas el 93 por ciento de las veces. La empresa cuenta con unos 40.000 usuarios. (Juergen, 2013)

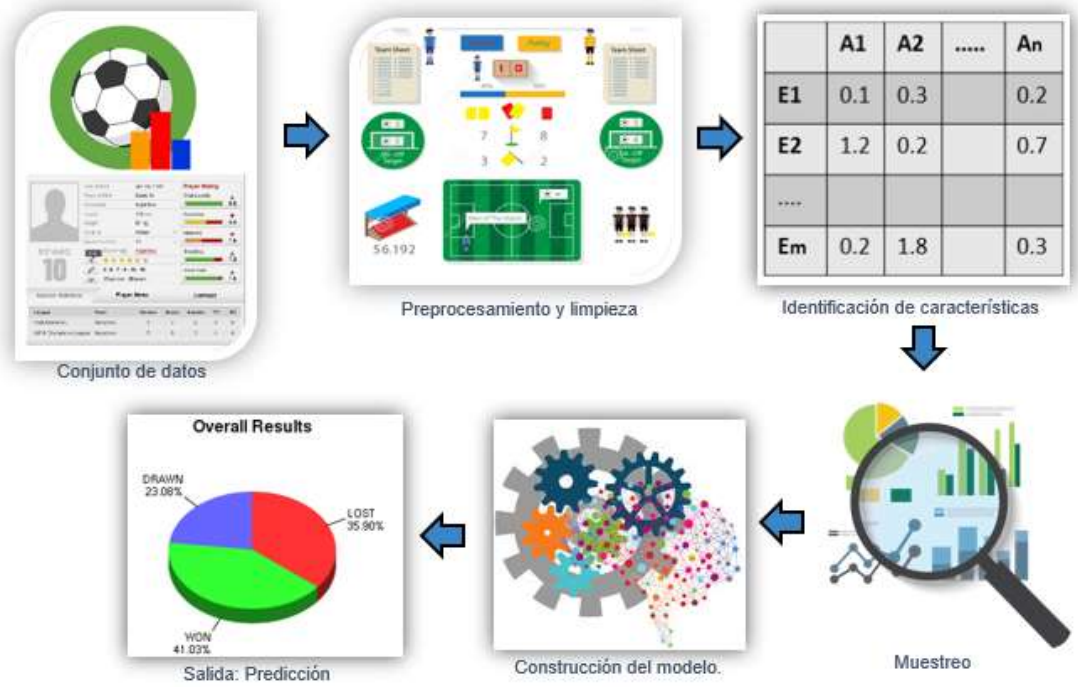
La compañía pretende utilizar modelos estadísticos y minería de datos para ayudar a los consumidores y las empresas a tomar decisiones más inteligentes en los mercados de deportes. Esto se realiza a través de la generación de fórmulas específicas y patentadas que analizan las estadísticas de una manera no convencional y científica. La compañía ha recibido una notable cobertura de prensa por haber elegido correctamente al ganador del Super Bowl XLV y haber superado con éxito a ESPN, Yahoo!, y CBS en los mercados de predicción de eventos deportivos.

1.8. METODOLOGÍA

Para poder desarrollar la metodología que seguirá el presente trabajo de investigación tecnológica fue necesario identificar el tipo de investigación el cual seguirá, por tanto, el paradigma de investigación que tendrá el presente trabajo será el positivista basado en el método científico, ya que se pretende dar un resultado objetivo con base en un enfoque centrado en el análisis de datos y la aplicación de algoritmos de aprendizaje de máquina que permitan desarrollar el experimento de software propuesto.

Para realizar el proceso de predicción del ganador de un partido de fútbol de la categoría A del fútbol profesional colombiano, es necesario realizar una serie de etapas en donde cada una se complementa directamente con la anterior, de esta forma se describen 6 pasos como se puede observar en la ilustración 18:

Ilustración 18. Metodología aplicada predicción fútbol.



Fuente: (Arias, 2019)

1.8.1 Conjunto de datos. Este hace referencia a la selección del subconjunto de todos los datos disponibles con los que trabajará. Se construye el conjunto de datos

con estadísticas y resultados de los partidos asociados a los equipos de la categoría A del fútbol profesional colombiano, comprendidos entre los años 2015 a la actualidad. Para poder construir este conjunto de datos es necesario implementar una herramienta de software que permita la extracción de información de forma eficiente utilizando web Scraping.[1]

"*Debe considerar qué datos necesita realmente para abordar el problema en el que está trabajando. Se debe realice algunas suposiciones sobre los datos que se necesitan.*" (Brownlee, 2013)

A continuación, hay se relacionan algunas preguntas para ayudar a pensar en este paso:

- ¿Cuál es el alcance de los datos que tiene disponibles?
- ¿Qué datos no están disponibles que se desearía tener disponible? Por ejemplo, datos que no están grabados o no pueden grabarse. Es posible que pueda derivar o simular estos datos.
- ¿Qué datos no necesitan para resolver el problema? Excluir datos es casi siempre más fácil que incluir datos.

1.8.2 Preprocesamiento y limpieza de datos. En esta etapa se realiza la selección de datos más menos importantes para el proceso de clasificación con el objetivo de conservar solo las características o los atributos más relevantes, se calcula la Matriz de dispersión para observar cuánto afecta un atributo a otro conjunto y sus correlaciones.

Una vez que haya seleccionado los datos, debe considerar cómo va a utilizar los datos. Este paso de preprocesamiento consiste en obtener los datos seleccionados en un formulario en el que pueda trabajar.[2]

1.8.3 Identificación de características. La literatura en el estado del arte sugiriere usar las características de si un equipo juega en casa o fuera de casa, estas características se transforman en una codificación en una variable discreta para su procesamiento posterior.

[1] *¿Qué es y cómo funciona Web Scraping?* Obtenido de: https://sitelabs.es/web-scraping-introduccion-y-herramientas/
[2] *¿Cómo hacer un buen procesamiento de información de aprendizaje de máquina?* Obtenido de: https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

Los métodos de selección de características ayudan a reducir las dimensiones sin perder mucho la información total. También ayuda a dar sentido a las características y su importancia. [1]

Existen 3 técnicas de identificación de características:

- Métodos de filtro. Métodos de filtro considera la relación entre las características y la variable de destino para calcular la importancia de las características.

- Métodos de envoltura. Los métodos de envoltura generan modelos con subconjuntos de características y miden el rendimiento de sus modelos.

- Métodos incrustados. La selección de funciones también se puede lograr con los conocimientos proporcionados por algunos modelos de aprendizaje automático.

1.8.4 Muestreo. En esta etapa se va a clasificar el conjunto de datos en dos de manera que una sea para realizar el entrenamiento y la segunda sea para realizar las pruebas.

Puede haber muchos más datos seleccionados disponibles de los que necesita para trabajar. Más datos pueden resultar en tiempos de ejecución mucho más largos para los algoritmos y mayores requisitos computacionales y de memoria. Se tomará una muestra representativa más pequeña de los datos seleccionados que pueden ser mucho más rápidos para explorar y crear prototipos de soluciones antes de considerar el conjunto de datos completo.[2]

1.8.5 Construcción modelo de predicción. En esta etapa, se aplica los clasificadores de aprendizaje automático necesarios para realizar predicción del ganador de un partido de fútbol de la categoría A del fútbol profesional colombiano.

1.8.6 Salida. En esta etapa se va a obtener el resultado que genera el modelo predictivo donde se indica el resultado del modelo predictivo.

---

[1] *¿Cómo preparar datos para el aprendizaje automático?* Obtenido de: https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

[2] *Técnicas de muestreo de aprendizaje de máquina*. Obtenido de: https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

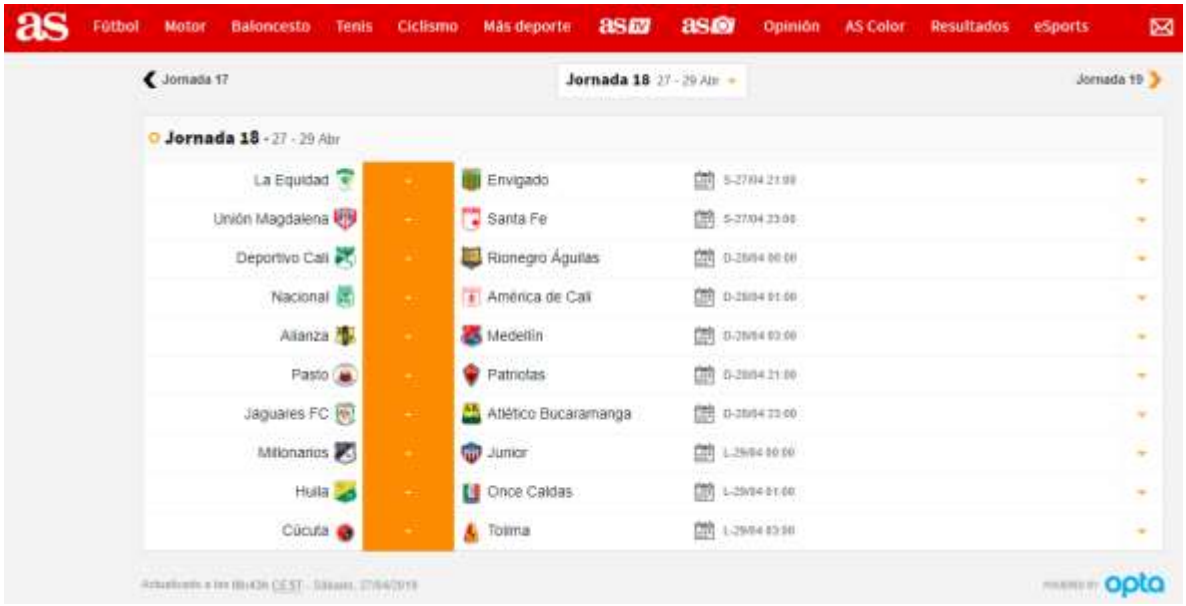
2. DESARROLLO DE LA METODOLOGÍA Y RESULTADOS

En esta sección se presentan los resultados obtenidos en cada una de las fases de la metodología propuesta.

2.1. CONJUNTO DE DATOS

Para construir el conjunto de datos, fue necesario indagar varias fuentes las cuales debían contar con la información relacionada a los partidos de fútbol de la categoría A del fútbol profesional colombiano, por otra parte, la información debía ser confiable en su totalidad, finalmente se optó por seleccionar la fuente de datos de la página as.com edición Colombia, el cual, es un periódico de origen español dedicado exclusivamente a los deportes siendo el fútbol el deporte principal. Otro factor de peso por el cual se seleccionó esta fuente de información es que sus widgets[1] son proporcionados por *Opta*[2] considerado como un proveedor mundial de datos deportivos detallados. En la ilustración 19 se puede ver la obtención de información de la jornada 18 de la liga águila 2019-1:

Ilustración 19. Obtención de información as.com Colombia.



[1] *¿Qué son los Widgets?* Obtenido de: https://neoattack.com/neowiki/widgets/
[2] *Widgets Fútbol Opta* Obtenido de: https://www.optasports.com/services/widgets/football/

Fuente: Colombia.as.com, (2019). Resultados fecha 19. Recuperado de: https://colombia.as.com/resultados/fútbol/colombia_i/jornada/

Luego de haber identificado la fuente de datos, fue necesario posteriormente identificar una herramienta que permitiera extraer la información de forma automática, es por esta razón, que selecciono la herramienta *Import.io*[1] la cual permite realizar integración de datos web de forma automática y permite descargar esta información en archivos de Excel.

Como resultado de lo anteriormente descrito, se identificaron 29 atributos, los cuales corresponden a los partidos de la categoría A del fútbol profesional colombiano comprendido entre los años 2015 a 2018, tanto el torneo de inicio como el torneo de finalización de la fase regular todos contra todos. Entre los datos obtenidos se tiene: la Etapa, la cual indica la fase en la cual se disputo el partido por ej. Fase regular (todos contra todos) semifinales, finales; el semestre; el año; la jornada; el partido (concatenación de los equipos local y visitante); equipo local; equipo visitante; puntos equipo local; puntos equipo visitante; posesión de pelota equipo local; posesión de pelota equipo visitante; tiros realizados equipo local; tiros realizados equipo visitante; faltas realizadas equipo local; faltas realizadas equipo visitante; fuera de lugar equipo local; fuera de lugar equipo visitante; tarjetas amarillas equipo local; tarjetas amarillas equipo visitante; día de la semana; día; mes; hora del partido; fecha del partido.

En la ilustración 20 se puede ver las variables del conjunto de datos y el tipo de dato:

[1] Integración de datos web Obtenido de: https://www.import.io/product/

Ilustración 20. Variables conjunto de datos.

Variable	Tipo Variable	Variable	Tipo Variable
Url	Cualitativa Nominal	Tiros al arco Local	Cuantitativa Discreta
Etapa	Cualitativa Ordinal	Tiros al arco Visitante	Cuantitativa Discreta
Semestre	Cualitativa Ordinal	Faltas Equipo Local	Cuantitativa Discreta
Año	Cuantitativa Discreta	Faltas Equipo Visitante	Cuantitativa Discreta
Fecha	Fecha	Fuera de juego Equipo Local	Cuantitativa Discreta
Partido	Cualitativa Nominal	Fuera de juego Equipo Visitante	Cuantitativa Discreta
Equipo Local	Cualitativa Nominal	Tarjetas Amarillas Equipo Local	Cuantitativa Discreta
Equipo Visitante	Cualitativa Nominal	Tarjetas Amarillas Equipo Visitante	Cuantitativa Discreta
FTR	Cualitativa Nominal	Tarjetas Rojas Equipo Local	Cuantitativa Discreta
Goles Equipo Local	Cuantitativa Discreta	Tarjetas Rojas Equipo Visitante	Cuantitativa Discreta
Goles Equipo Visitante	Cuantitativa Discreta	Dia de la semana	Cuantitativa Discreta
Puntos Local	Cuantitativa Discreta	Día del mes	Cualitativa Ordinal
Puntos Visitante	Cuantitativa Discreta	Mes	Cualitativa Ordinal
Posesion Local	Cuantitativa Discreta	Hora	Fecha
Posesion Visitante	Cuantitativa Discreta	Fecha partido	Fecha

(Arias, 2019)

A continuación, en la ilustración 21 se da un resumen de la información de extracción de información:

Ilustración 21. Resultados fase construcción conjunto de datos.

(Arias, 2019)

2.2. PREPROCESAMIENTO Y LIMPIENZA DE INFORMACIÓN

En esta fase se procede a eliminar y transformar la información. Para el conjunto de datos obtenido de la fase anterior, fue necesario construir la variable FTR (*full time result* – Resultado tiempo completo) en base a los puntos obtenidos por el equipo local y visitante de la siguiente manera como se observa en la ilustración 22:

Ilustración 22. Construcción variable dependiente.

Puntos Local	Puntos Visitante	FTR (Resultado tiempo final)
3	0	L
1	1	E
0	3	V

(Arias, 2019)

Dicha variable es de tipo cualitativa nominal, el cual contiene la información del ganador de dicho partido categorizado de la siguiente manera: 'L', el cual indica si el equipo que jugo en condición de local fue el ganador del partido; 'V', el cual indica si el equipo que jugo en condición de visitante fue el ganador del partido y finalmente 'E', el cual indica si hubo un empate. Dicha variable tiene una función muy importante dentro del conjunto de datos ya que será la variable predictora del conjunto de datos.

Por otra parte, no se toman los siguientes atributos ya que no al modelo de predicción: Url, Etapa, DiaNom, Dia, Mes, Hora, Fec_Partido, Sem, Anno, Fecha, Partido. En cuanto a las variables que contienen los nombres de los equipos de fútbol, fue necesario cambiarlos para que todos quedaran estandarizados, es decir, con el mismo nombre, ya que se requeriré en fases posteriores para actualizar la información. Posteriormente, se eliminan aquellos registros donde el equipo Unión Magdalena figura como equipo local o visitante, dado que dicho equipo estuvo en la categoría B del fútbol profesional colombiano para el intervalo de años el cual está delimitado en el presente trabajo de investigación tecnológica.

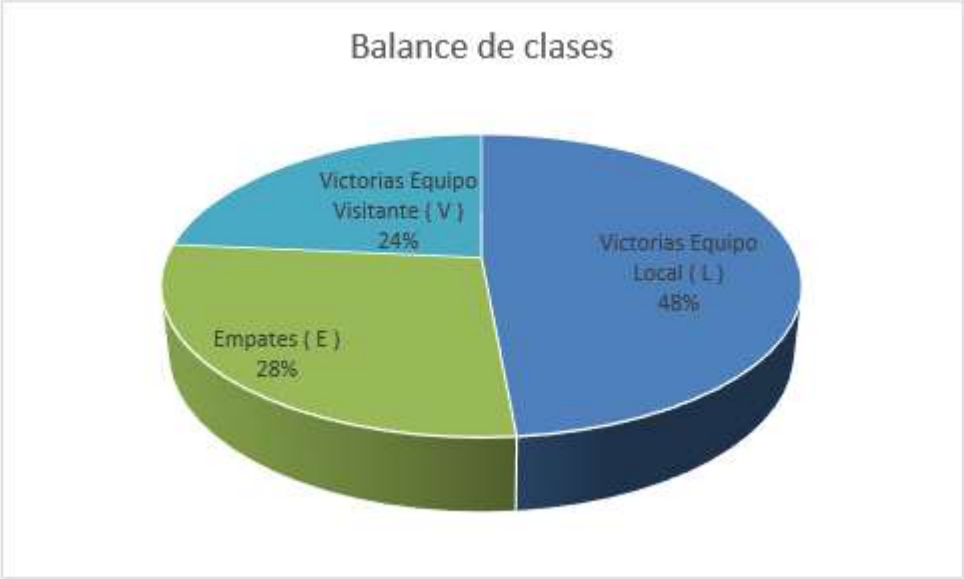
Finalmente, se valida la información en busca de datos incongruentes o faltantes, como resultado se identificaron 6 registros los cuales no tenían relacionado el porcentaje posesión del balón, por lo tanto, fue necesario realizar la actualización de la información de forma manual. En cuanto a datos atípicos no se identificó alguno.

2.3. IDENTIFICACIÓN Y SELECCIÓN DE CARACTERISTICAS

En esta fase se seleccionan aquellas características (atributos) los cuales serán tomados en cuenta para construcción del modelo de predicción, por tanto, se realizarán análisis univariados y bivariados para poder establecer relaciones de asociación o relaciones simples entre la variable predictora y las variables independientes.

La primera variable en ser analizada es la variable predictora FTR (Resultado tiempo final), contiene las características de: gana el equipo local, gana el equipo visitante o hay un empate, estas clases son mutuamente excluyentes. En la ilustración 23 se puede ver la proporción de balanceo de las clases correspondientes a los años 2015 a 2018:

Ilustración 23. Balance clases conjunto de datos.



(Arias, 2019)

En base a la anterior ilustración, se puede notar que no existe un balanceo de clases debido a la naturaleza de los datos que proporciona la categoría A del fútbol profesional colombiano, los datos desequilibrados generalmente se refieren a un problema de clasificación donde las clases no están representadas por igual, pero este desequilibrio de clases no solo es común, sino que se espera, ya que

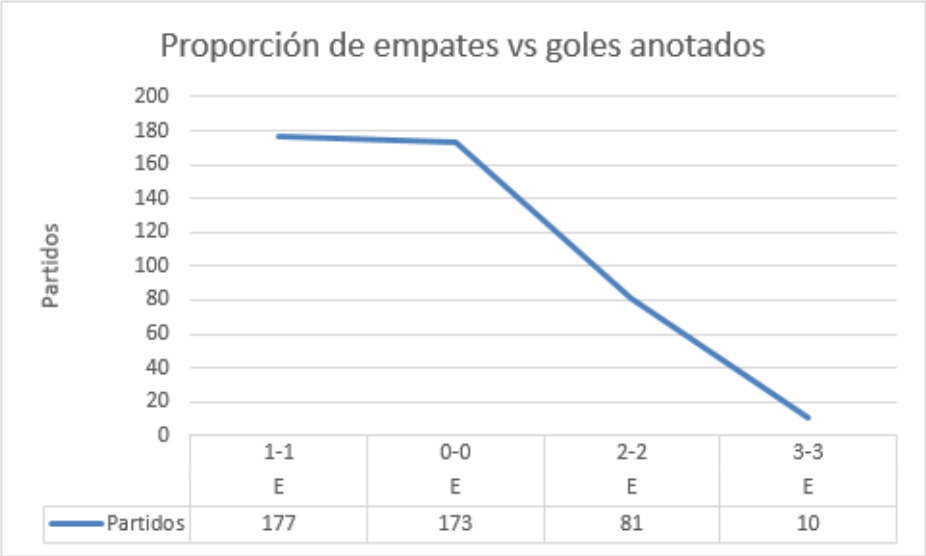
normalmente los equipos de fútbol colombiano suelen defender su localía debido a factores como el apoyo de su afición. En la sección 2.4 del presente documento se parametrizan los algoritmos de aprendizaje de máquina para conjuntos de datos no balanceados.

Posteriormente, se realiza un análisis bivariado, entre las variables FTR, y los goles, ya que es una variable que está directamente relacionada, ya que se requiere anotar goles para poder ganar, empatar o perder un partido de fútbol.

Ilustración 24. Proporción de victorias equipo local, vs goles anotados.

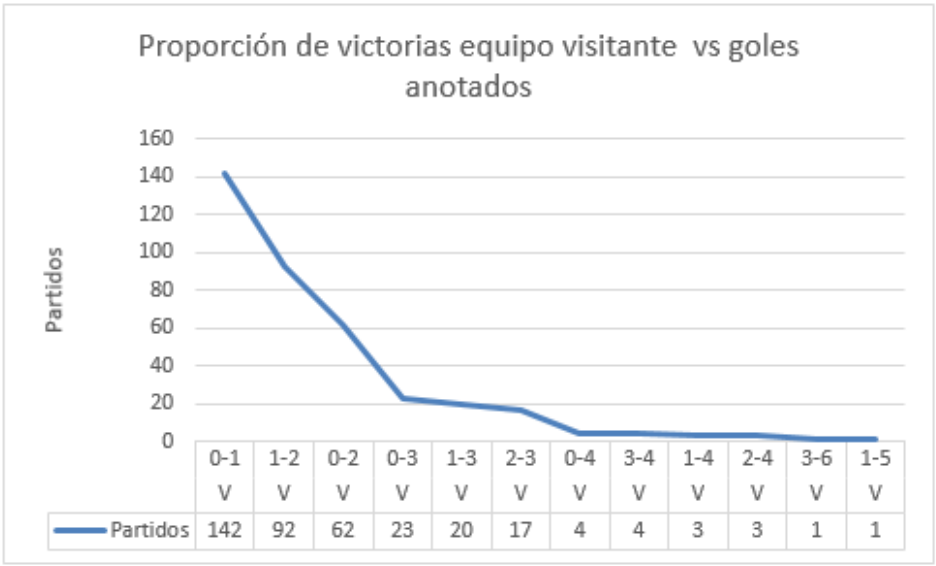
(Arias, 2019)

Ilustración 25. Proporción de empates vs goles anotados.



(Arias, 2019)

Ilustración 26. Proporción de victorias equipo visitante vs goles.

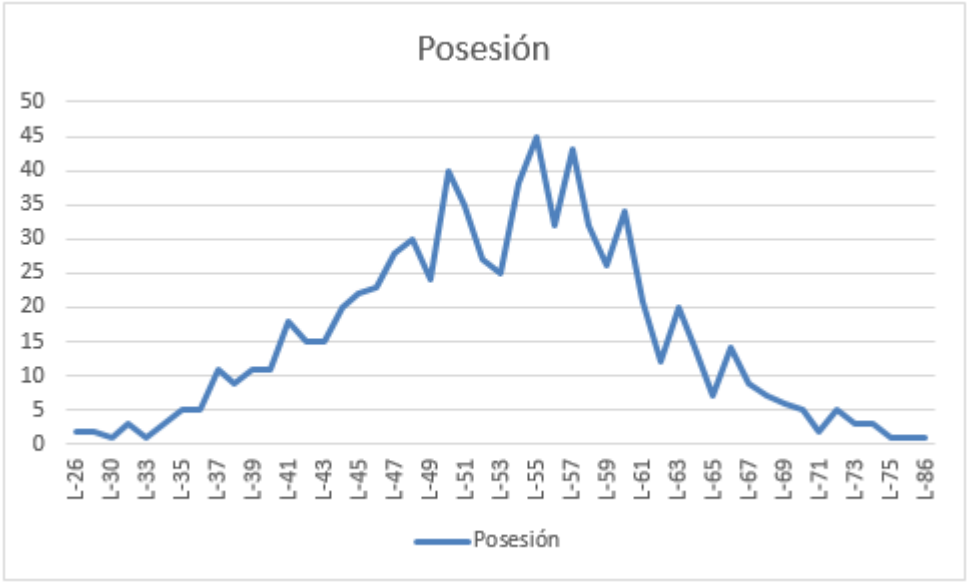


(Arias, 2019)

Como se puede ver en las ilustraciones 24, 25 y 26, por cada clase de la variable predictora, se agrupo los marcadores más frecuentes que suele haber en el fútbol colombiano, teniendo como, por ejemplo, que el marcador más frecuente porque el que gana un equipo local es 1 a 0 con 235 partidos que corresponden al 14,87% del total de datos, el marcador más frecuente en empate es 1 a 1 con 177 partidos, y el marcador por el cual un equipo visitante más gana es 0 a 1 con 142 partidos. En base a la asociación encontrada entre los goles y la variable predictora, las variables goles de local y de visitante se tomarán en cuenta para la construcción final del conjunto de datos.

El siguiente análisis, corresponde a la posesión del balón con el resultado de tiempo final (FTR), como se muestra en la ilustración 27, la posesión del balón no es un atributo que logre explicar a profundidad que un equipo en condición de local gane un partido, como se puede ver, existe posesiones de balón de equipos locales inferiores al 50%, es decir, durante el transcurso del partido predominó el equipo visitante con la posesión del balón, pero esto no quiere decir que el equipo visitante haya ganado el encuentro, por el contrario, el equipo local salió victorioso a pesar de la baja posesión del balón.

Ilustración 27. Análisis posesión de balón.



(Arias, 2019)

El análisis a continuación corresponde a los tiros al arco realizados por los equipos en condición de local, este atributo está relacionado con los goles marcados, ya que cada uno de goles cuenta como un tiro al arco, por ende, también está asociado e influye en el ganador de un partido de fútbol, al revisar la ilustración 28 se puede ver que la cantidad de tiros al arco realizados por equipos locales es de 5 tiros en 143 partidos, por otra parte, rara vez se puede ocurrir que no se realicen tiros al arco, en dicha ocasión con 4 partidos de acuerdo a la información recolectada en el conjunto de datos. Por tanto, dicha variable será tenida en cuenta para la construcción del conjunto de datos final.

Ilustración 28. Análisis tiros al arco.

(Arias, 2019)

En base a los análisis previos, se toma como base los goles marcados por los equipos de fútbol tanto de local y visitante y se deduce las siguientes variables[1]: fuerza de ataque como local, fuerza de defensa como local, fuerza de ataque como visitante, fuerza de defensa como visitante, dichas variables se describen en la ilustración 29:

[1] *Define variables: attack & defence strength* Obtenido de: https://beatthebookie.blog/2017/05/16/define-variables-attack-defence-strength/

Ilustración 29. Ecuaciones fuerza de ataque y defensa.

$$Fuerza\ Ataque\ Local = \frac{Promedio\ Goles\ Marcados\ Condición\ Local}{Promedio\ Total\ Goles\ Marcados\ Condición\ Local}$$

$$Fuerza\ Defensa\ Local = \frac{Promedio\ Goles\ Recibidos\ Condición\ Local}{Promedio\ Total\ Goles\ Recibidos\ Condición\ Local}$$

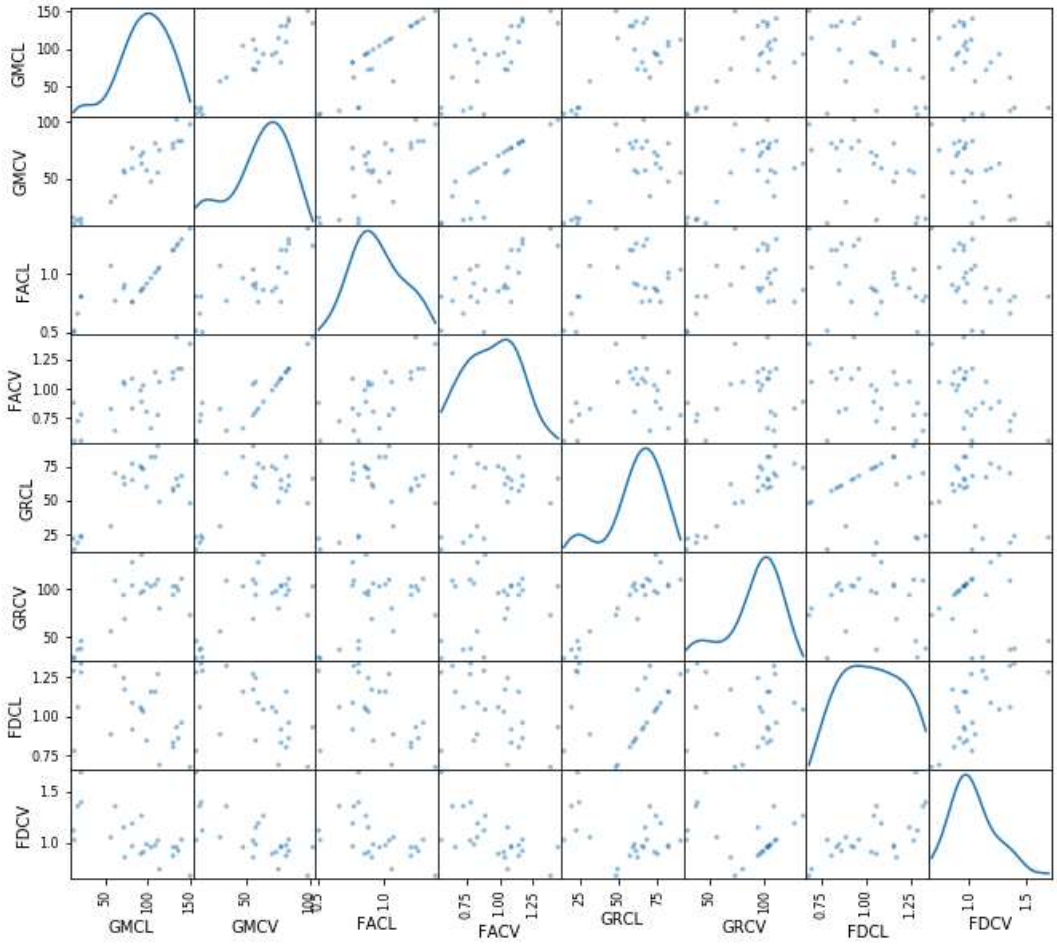
$$Fuerza\ Ataque\ Visitante = \frac{Promedio\ Goles\ Marcados\ Condición\ Visitante}{Promedio\ Total\ Goles\ Marcados\ Condición\ Visitante}$$

$$Fuerza\ Defensa\ Visitante = \frac{Promedio\ Goles\ Recibidos\ Condición\ Visitante}{Promedio\ Total\ Goles\ Recibidos\ Condición\ Visitante}$$

(Arias, 2019)

Posteriormente se agregar otras variables al conjunto de datos, como la cantidad de goles marcados como local, goles marcados como visitante, Goles recibidos como local, y goles recibidos como visitante. Después de agregar los nuevos datos, como se puede observar en la ilustración 30 se procede a realizar un análisis de asociación identificando así, relación entre los goles marcados tanto de local y visitante como las fuerzas de ataque en condición de local y visitante respectivamente.

Ilustración 30. Análisis de correlación de datos.



(Arias, 2019)

Al finalizar los datos agregados a cada uno de los equipos queda de la siguiente forma como se observa en la tabla 1:

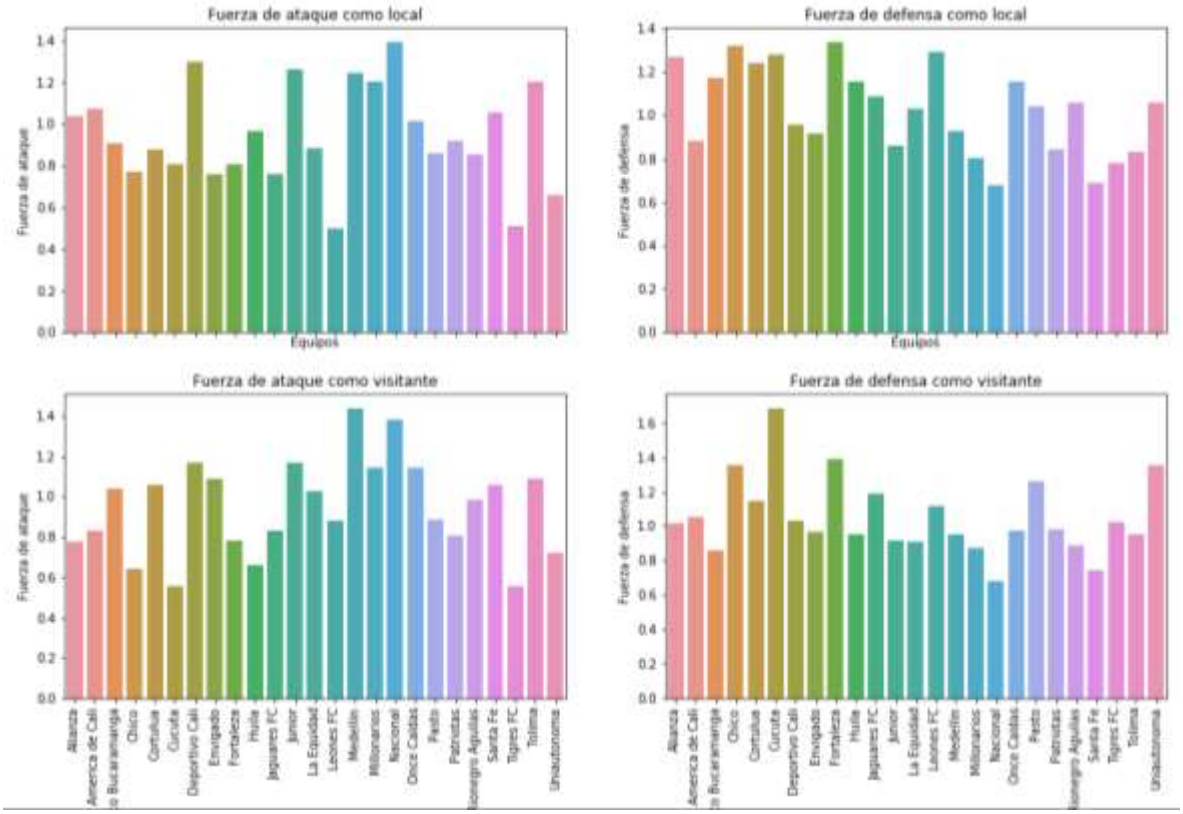
Tabla 1. Tabla de características previas.

	TEAM	GMCL	GMCV	FACL	FACV	GRCL	GRCV	FDCL	FDCV
0	Alianza	112	55	1.039443	0.775740	90	110	1.269394	1.020882
1	America de Cali	57	29	1.071569	0.828541	31	56	0.885682	1.052769
2	Atletico Bucaramanga	73	55	0.907153	1.038703	62	69	1.170902	0.857446
3	Chico	62	34	0.770459	0.642108	70	109	1.321986	1.354516
4	Cortulua	72	57	0.879814	1.058533	67	94	1.244241	1.148647
5	Cucuta	22	10	0.806497	0.557123	23	46	1.281382	1.686311
6	Deportivo Cali	140	83	1.299304	1.170663	68	111	0.959097	1.030162
7	Envigado	82	77	0.761021	1.086037	65	104	0.916784	0.965197
8	Fortaleza	22	14	0.806497	0.779972	24	38	1.337094	1.393039
9	Huila	104	47	0.965197	0.662906	82	103	1.156559	0.955916
10	Jaguares FC	82	59	0.761021	0.832158	77	128	1.086037	1.187935
11	Junior	136	83	1.262181	1.170663	61	99	0.860367	0.918794
12	La Equidad	95	73	0.881671	1.029619	73	98	1.029619	0.909513
13	Leones FC	13	15	0.501649	0.879667	22	29	1.290179	1.119062
14	Medellin	134	102	1.243619	1.438646	66	103	0.930889	0.955916
15	Millonarios	130	81	1.206497	1.142454	57	94	0.803949	0.872390
16	Nacional	150	98	1.392111	1.382228	48	73	0.677010	0.677494
17	Once Caldas	109	81	1.011601	1.142454	82	105	1.156559	0.974478
18	Pasto	93	63	0.863109	0.888575	74	136	1.043724	1.262181
19	Patriotas	99	57	0.918794	0.803949	60	106	0.846262	0.983759
20	Rionegro Aguilas	92	70	0.853828	0.987306	75	96	1.057828	0.890951
21	Santa Fe	114	75	1.058005	1.057828	49	80	0.691114	0.742459
22	Tigres FC	14	10	0.513225	0.557123	14	28	0.779972	1.026450
23	Tolima	130	77	1.206497	1.086037	59	103	0.832158	0.955916
24	Uniautonoma	18	13	0.659861	0.724260	19	37	1.058533	1.356381

(Arias, 2019)

Posteriormente, en la ilustración 31 se observa las fuerzas de ataque y defensa en condición de local y visitante, para ver cómo se comporta cada uno de los equipos de fútbol de la categoría A del fútbol profesional colombiano.

Ilustración 31. Fuerza de ataque y defensa equipos del FPC.



(Arias, 2019)

Al verificar la ilustración 31, se puede observar que en condición de local (Las primeras dos imágenes de arriba hacia abajo), el equipo que presenta mejores números, es decir, que marca muchos goles como local y recibe pocos en la misma condición, es el equipo Atlético Nacional, de forma contraria, el equipo con más bajo rendimiento es el equipo Leones FC, posiblemente esta sea una de las razones por la cual el equipo este de nuevo en segunda división del fútbol profesional colombiano[1]. Al validar la información como visitante (dos últimos gráficos de arriba abajo respectivamente) se puede ver que el equipo Independiente Medellín es uno de los mejores atacantes en condición de visitante, al igual que su equipo de patio Atlético Nacional, en cuanto a defensa, el equipo Cúcuta Deportivo es uno de los equipos a los cuales más goles le marcan en condición de visitante seguido del equipo Fortaleza FC.

[1] *Leones, sin pena ni gloria, regresa a la segunda división* Obtenido de: https://caracol.com.co/radio/2018/10/14/deportes/1539531247_689815.html

Posteriormente, se procede a agregar el comportamiento de goles en las últimas 2 temporadas, para esto se calcula el promedio de goles marcados y recibidos y el promedio de tiros al arco en condición de local y visitante de los últimos 38 partidos, esto se hace teniendo en cuenta que los equipo pueden bajar o subir la tendencia de goles marcados y tiros al arco partido tras partido.

Como resultado, se obtiene los datos los cuales será utilizados por los algoritmos de predicción que contienen los atributos: FTR (Resultado tiempo completo) 'L' Gana el equipo local, 'E' si empatan los dos equipos y 'V' si gana el equipo visitante; FTR será la variable dependiente de los algoritmos de aprendizaje de máquina, es decir, la variable a predecir. Y como variables independientes de obtiene: Fuerza de ataque como local, fuerza en defensa como local, fuerza de ataque como visitante, fuerza de defensa como visitante, goles marcados local, goles marcados como visitante, tiros al arco como local y tiros al arco como visitante.

A continuación, en la tabla 2, se muestra el conjunto de datos el cual será procesado por los algoritmos de aprendizaje de máquina.

Tabla 2. Variables modelo de predicción.

	FTR	FACL	FDCL	FACV	FDCV	utGolesLoc	utGolesVis	utTirosLoc	utTirosVis
1200	L	1.243619	0.930889	0.662906	0.955916	1.500000	0.947368	4.842105	3.684211
1201	E	0.853828	1.057828	0.888575	1.262181	0.710526	1.289474	3.631579	4.263158
1202	V	0.501649	1.290179	0.828541	1.052769	0.026316	1.157895	0.026316	4.000000
1203	L	1.011601	1.156559	0.832158	1.187935	0.947368	1.052632	3.421053	3.763158
1204	E	1.299304	0.959097	1.086037	0.965197	1.289474	0.868421	4.289474	3.736842
1205	L	1.262181	0.860367	1.038703	0.857446	1.394737	0.763158	5.026316	3.815789
1206	V	1.206497	0.832158	1.382228	0.677494	1.157895	1.342105	3.710526	4.263158
1207	V	1.058005	0.691114	0.803949	0.983759	0.947368	0.947368	3.973684	3.421053
1208	E	0.770459	1.321986	1.142454	0.872390	0.894737	1.368421	3.105263	5.000000
1209	E	0.881671	1.029619	0.775740	1.020882	0.973684	0.973684	3.947368	3.421053
1210	L	1.299304	0.959097	0.879667	1.119062	1.263158	0.052632	4.105263	0.157895
1211	L	0.853828	1.057828	1.029619	0.909513	0.763158	0.973684	3.763158	3.868421
1212	L	0.770459	1.321986	1.057828	0.742459	0.947368	0.947368	3.184211	3.894737
1213	E	1.243619	0.930889	1.086037	0.955916	1.552632	1.210526	4.763158	3.736842
1214	L	1.262181	0.860367	1.142454	0.974478	1.473684	0.947368	5.105263	3.394737
1215	L	0.863109	1.043724	0.828541	1.052769	1.263158	1.157895	4.289474	3.973684
1216	L	0.907153	1.170902	0.775740	1.020882	0.842105	1.000000	3.868421	3.500000

(Arias, 2019)

Como último paso se realiza un análisis de correlación de datos al conjunto de variables independientes, en la ilustración 32 se muestra los resultados:

Ilustración 32. Análisis de corrección de variables.

	FACL	FDCL	FACV	FDCV	utGolesLoc	utGolesVis	utTirosLoc	utTirosVis
FACL	1.000000	-0.544332	-0.022247	0.017502	0.555961	-0.014244	0.431274	-0.010061
FDCL	-0.544332	1.000000	0.016672	-0.015511	-0.347343	0.004249	-0.310915	-0.001752
FACV	-0.022247	0.016672	1.000000	-0.600266	-0.005778	0.513627	0.011111	0.405100
FDCV	0.017502	-0.015511	-0.600266	1.000000	-0.038297	-0.385275	-0.058916	-0.333507
utGolesLoc	0.555961	-0.347343	-0.005778	-0.038297	1.000000	0.436238	0.932249	0.483695
utGolesVis	-0.014244	0.004249	0.513627	-0.385275	0.436238	1.000000	0.490569	0.932416
utTirosLoc	0.431274	-0.310915	0.011111	-0.058916	0.932249	0.490569	1.000000	0.538523
utTirosVis	-0.010061	-0.001752	0.405100	-0.333507	0.483695	0.932416	0.538523	1.000000

(Arias, 2019)

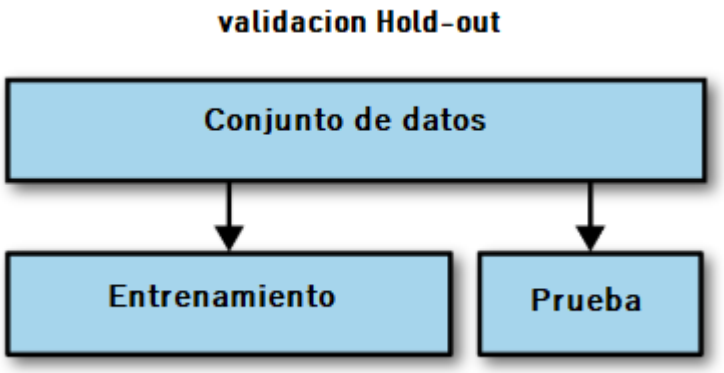
En base a la ilustración 32, se puede ver que existen diversas correlaciones entre las variables, entre las más destacadas se tiene: que a mayor fuerza de ataque como local (FACL) hay menos fuerza de defesa como local (FDCL), por el contrario, no existe correlación entre las fuerzas de ataque en condición de local y visitante ya que el valor de correlación está próximo a cero, es por ende que los equipos del fútbol colombiano son mejores jugando en condición de local, que jugando en condición de visitante. Por otra parte, existe correlación entre la fuerza de ataque y los tiros al marco realizados en las dos últimas temporadas y existe una correlación positiva fuerte entre los goles marcados en las últimas temporadas con los tiros realizados, estos atributos, de acuerdo con el análisis de correlación pueden explicar su comportamiento entre sí, y explicar que los cambios en algunas de ellas pueden llegar a explicar el comportamiento de otros atributos.

2.4. MUESTREO

En esta fase, se procede a realizar lo que en aprendizaje de máquina se denomina *data Splitting*, o dicho en español división de datos. Consiste en términos prácticos en separar el conjunto de datos en un subconjunto que servirá para entrenar el modelo de predicción y otro subconjunto que servirá como prueba.

El método que se empleó en el presente trabajo de investigación tecnológica es el *Holdout* o método de retención el cual se puede observar en la ilustración 33, consiste en definir a partir de un conjunto de datos inicial, un conjunto de entrenamiento y otro conjunto de prueba, Se establece una mayor proporción de datos para entrenamiento y los datos de prueba el restante. Se escogió este método ya que a la hora de computar es rápido, también debido a su sencillez, por otra parte, se requiere predecir los partidos más actuales en base a los datos ya jugados, por ende, los datos de la temporada 2018-1 y 2018-2 serán tomados como prueba, mientras los datos del 2015 a 2017 será tomados como entrenamiento.

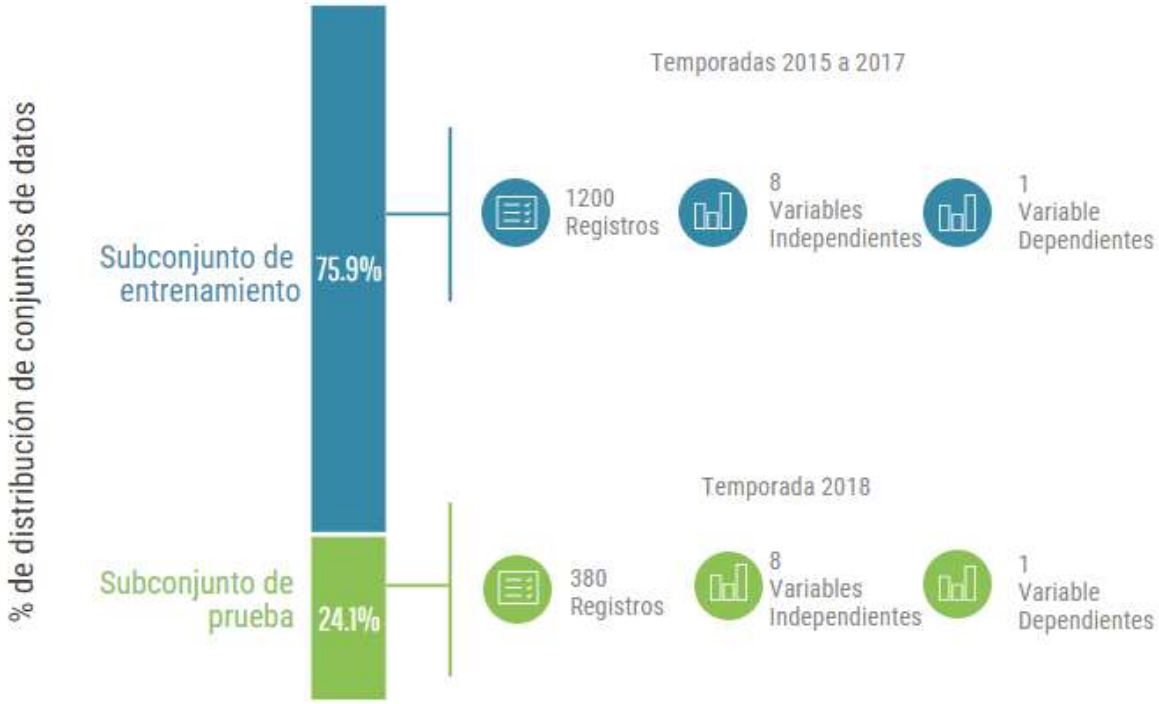
Ilustración 33. Método *Hold-out*.



(Arias, 2019)

En las ilustraciones 34 y 35 se describe el muestreo del conjunto de datos:

Ilustración 34. Distribución del conjunto de datos.



(Arias, 2019)

Ilustración 35. Características conjunto entrenamiento datos 2019.

(Arias, 2019)

En esta fase de la metodología, en la fase de muestreo es requerido y de importancia verificar que las clases estén balanceadas de forma equilibrada, es decir, que cada una de las clases tenga la misma cantidad de registros. Para el caso del conjunto de datos del fútbol profesional colombiano se puede ver en la ilustración

23, las clases la variable predictora no siguen un balanceo, siendo la clase mayoritaria victoria equipo local (L).

De acuerdo a la investigación sobre los diversos métodos de balanceo existentes, se encuentra el muestreo estratificado, el cual consiste en construir un conjunto de entrenamiento y prueba de forma balanceado dando como parámetros los porcentajes de distribución de cada conjunto de datos (entrenamiento y prueba) y el arreglo de clases, el inconveniente es que el conjunto de datos desde su construcción ya debe estar balanceado, por tanto, al aplicar este método a un conjunto de datos no balanceado, los porcentajes de distribución de los conjunto de datos estarán igualmente desbalanceados.

Al aplica el método de submuestreo el cual consiste en eliminar muestras de clases sobre representadas, para este caso, victorias para el equipo local, provocará que el conjunto de datos pierda datos importantes que afecten los resultados finales al tener que eliminar partidos en los cuales hayan ganado en condición de local, pudiendo afectar el desempeño de equipos que regularmente son buenos en dicha condición, como lo es el caso de Atlético nacional.

De forma contraria, se sugiere utilizar el método de sobre muestreo, el cual consisten en agregar más datos de las clases subrepresentadas, para este caso empates y victorias del equipo en condición de visitante, al igual que el submuestreo, se corre el riesgo de modificar y alterar los valores originales de los partidos del fútbol profesional colombiano al tener que agregar muestras sintéticas al conjunto de datos[1].

Después de consultar varias estrategias de balanceo de clases, se optó por seleccionar la implementación del hiperparametro *class_weight,* el cual está presente en los algoritmos de aprendizaje de máquina implementados en el presente trabajo de investigación tecnológica.

Muchos algoritmos de aprendizaje de máquina en *scikit-learn* viene con un método incorporado para manejar clases desequilibradas. Si se tiene clases altamente desequilibradas, existe la opción de usar el parámetro *class_weight* para ponderar las clases para asegurar que se tiene una combinación de registros equilibrada de cada clase. Específicamente, el argumento '*balanced'* pesará automáticamente las

[1] *clasificación multiclase con conjunto de datos desequilibrados* Obtenido de: https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a

clases de forma inversamente proporcional a su frecuencia. En la ilustración 36 se muestra el algoritmo de balanceo de clases:

Ilustración 36. Ecuación de balanceo de clases.

$$w_j = \frac{n}{k\, n_j}$$

Fuente: chrisalbon.com, (2017). Manejo de clases desequilibradas. Recuperado de: https://chrisalbon.com/machine_learning/logistic_regression/handling_imbalanced_classes_in_logistic_regression/

donde Wj es el peso para la clase j, n es el número de observaciones, nj es el número de observaciones en la clase j y k es el número total de clases.

2.5. CONSTRUCCIÓN MODELO PREDICTIVO

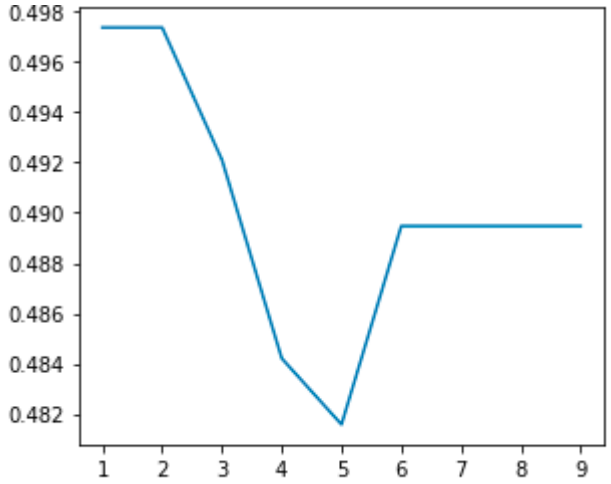
En base al muestreo del subconjunto de datos, se procede a implementar algunos de los algoritmos de aprendizaje de máquina que mejores resultados obtuvieron de acuerdo con la literatura en el estado del arte. Los algoritmos que se seleccionaron fueron: regresión logística, *Random Forest* y Máquina de soporte vectorial.

2.5.1 Regresión Logística multiclase. Para implementar el algoritmo de regresión logística se tuvieron en cuenta algunos parámetros en la implementación de la librería *Scikit Learn* de Python, entre ellos: solver=' lbfgs' y multi_class='multinomial', los cuales se ajustan a problemas multiclase, como lo es el caso de la predicción del ganador de un partido de la categoría A del fútbol profesional colombiano el cual tiene 3 clases (local, empate, visitante). Por otra parte, se valida el parámetro '*C*' el cual es denotado como la inversa de la fuerza de regularización, consiste en "*aplicar una penalización al aumentar la magnitud de los valores de los parámetros para reducir el sobreajuste. Cuando se entrena un modelo como un modelo de regresión logística, se está eligiendo parámetros que le dan el mejor ajuste a los datos. Esto significa minimizar el error entre lo que el modelo predice para su variable dependiente, dados sus datos en comparación con lo que realmente es su variable dependiente*" … (code.i-harness.com, 2019)

Dado lo anterior, se debía buscar el valor del parámetro '*C*' que mejor predicción tuviera para el algoritmo de regresión logística multiclase, por tanto, se realiza una iteración sobre un arreglo de variables cuantitativas continuas y discretas positivas (mayores a cero) con el objetivo de identificar el valor de '*C*' que tuviera mejor puntuación de precisión, es decir, mayor cantidad de predicciones correctas sobre cada una de las clases.

Al ejecutar el algoritmo e iterar sobre cada uno de los valores del arreglo, para definir el valor del parámetro 'C', en la ilustración 37 que muestra la precisión del algoritmo:

Ilustración 37. Exactitud en función de la constante de penalización.

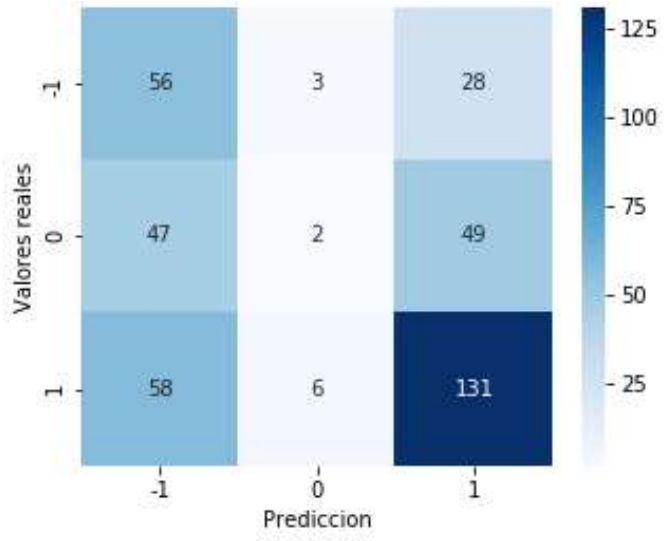


(Arias, 2019)

Como se ve en el gráfico anterior, el algoritmo tuvo una mejor precisión con un valor del parámetro '*C*'=0, los valores superiores a 0 desmejoraban su puntuación de precisión, por tanto, se definió este valor para dicho parámetro.

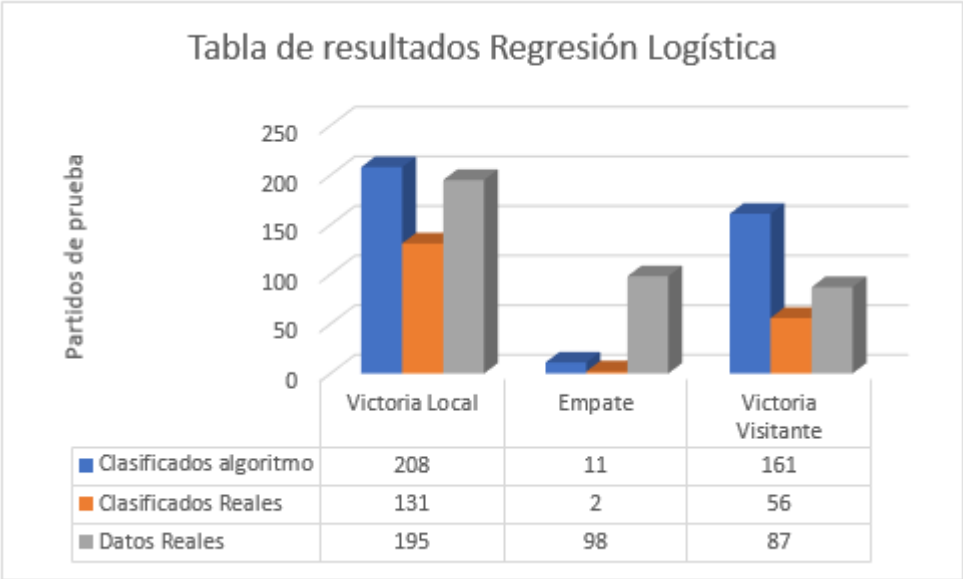
Una vez identificado el valor de '*C*' se valida la precisión del algoritmo con la métrica matriz de con función, en la ilustración 38 se muestra los resultados de la matriz de confusión, en la tabla 3 se describe dicha matriz de confusión y el reporte de resultados se observa en la ilustración 39:

Ilustración 38. Matriz de confusión regresión logística multiclase.



(Arias, 2019)

Tabla 3. Tabla de resultados regresión logística multiclase.



(Arias, 2019)

Ilustración 39. Reporte de resultados regresión logística multiclase.

accuracy score: 0.49736842105263157

	precision	recall	f1-score	support
Victoria Visitante	0.34	0.61	0.44	87
Empate	0.36	0.19	0.25	98
Victoria Local	0.66	0.58	0.62	195
micro avg	0.49	0.49	0.49	380
macro avg	0.45	0.46	0.44	380
weighted avg	0.51	0.49	0.48	380

(Arias, 2019)

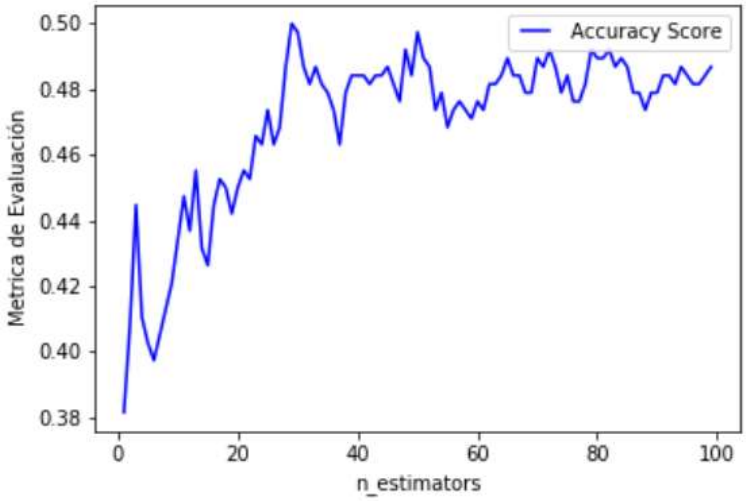
De acuerdo con la matriz de confusión, el algoritmo tuvo una puntuación de precisión del 0,4973, en cuanto a la exactitud (casos en que el algoritmo predijo correctamente la clase i de todas las instancias en las que el algoritmo predijo i) por cada una de las clases, el algoritmo se ajusta mejor a predecir los que ganan en condición de local y visitante, en cuanto a partidos empatados fueron pocos los que el algoritmo predijo. El *recall* (casos en los que el algoritmo predijo correctamente i de todos los casos que están etiquetados como i), se obtuvo un buen resultado en la predicción de casos donde gana un equipo visitante con un 0.61, seguido de un 0,58 para equipos que ganan en condición de local y en último lugar con 0,19 de poder de predicción en cuanto empates se refiere.

2.5.2 Random Forest. Para implementar el algoritmo *Random Forest*, se tuvieron en cuenta algunos parámetros en la implementación de la librería *Scikit Learn* de Python, entre ellos *n_estimators*, el cual representa el número de árboles en el bosque. Por lo general, cuanto mayor es el número de árboles, mejor aprende el algoritmo de los datos. Otro parámetro para tener en cuenta fue *max_depth*, representa la profundidad de cada árbol en el bosque. Cuanto más profundo es el árbol, más divisiones tiene y captura más información sobre los datos.

Para poder encontrar los valores para *n_estimators* y *max_depth*, se realiza una serie de ciclos anidados que permitieran encontrar el punto óptimo en donde la puntuación de precisión fuera el mejor, al finalizar el ciclo de iteraciones, se identificó los valores para *n_estimators* y *max_depth* siendo 29 y 14 respectivamente.

En la ilustración 40 muestra el comportamiento de las iteraciones del parámetro *n_estimators* el cual tuvo una puntuación de precisión máxima del 0. 5:

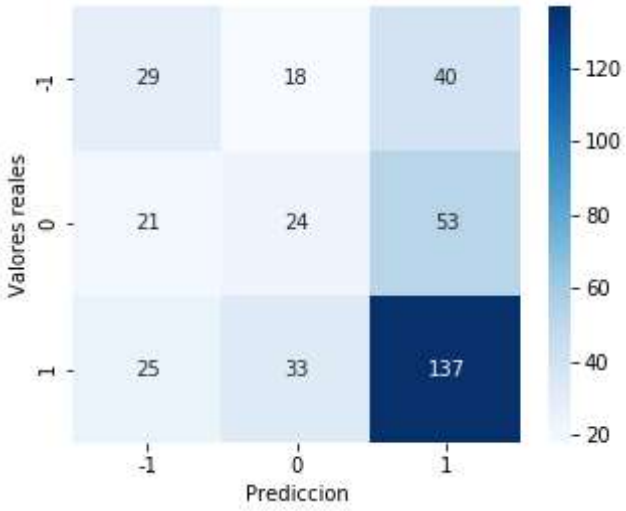
Ilustración 40. Precisión en función del parámetro *n_estimator*.



(Arias, 2019)

Posteriormente se valida el algoritmo con la métrica de precisión matriz de confusión obteniendo así los siguientes resultados:

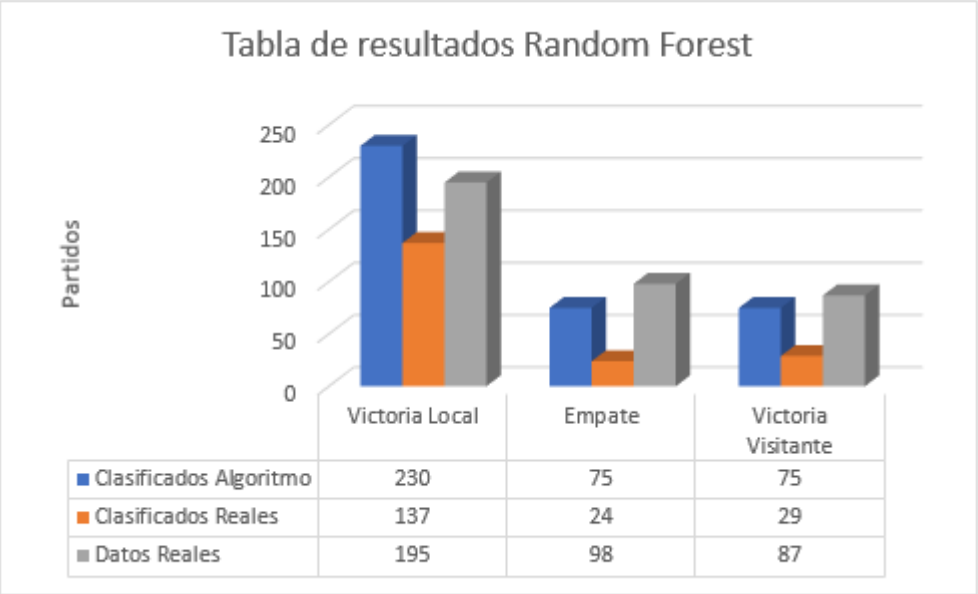
Ilustración 41. Matriz de confusión *random forest*.



(Arias, 2019)

De acuerdo con la ilustración 41, la matriz de confusión indica que el algoritmo *Random Forest* categorizo 230 partidos como victoria de equipos locales, siendo categorizados de forma correcta 137 partidos de 195 partidos reales ganados por equipos locales. Posteriormente el algoritmo categorizó 75 partidos ganados por equipos visitantes, siendo categorizados de forma correcta 29 de 87 partidos reales ganados por equipos visitantes y finalmente 75 partidos categorizados como empate, categorizando de forma correcta 24 de 98 empates reales según el conjunto de datos de prueba, los anterior, se puede visualizar gráficamente en la tabla 4:

Tabla 4. Tabla de resultados *random forest*.



(Arias, 2019)

Ilustración 42. Reporte de resultados random forest.

```
accuracy score:  0.5

                precision    recall  f1-score   support

Victoria Visitante    0.39      0.33      0.36        87
Empate                0.32      0.24      0.28        98
Victoria Local        0.60      0.70      0.64       195

   micro avg          0.50      0.50      0.50       380
   macro avg          0.43      0.43      0.43       380
weighted avg          0.48      0.50      0.48       380
```

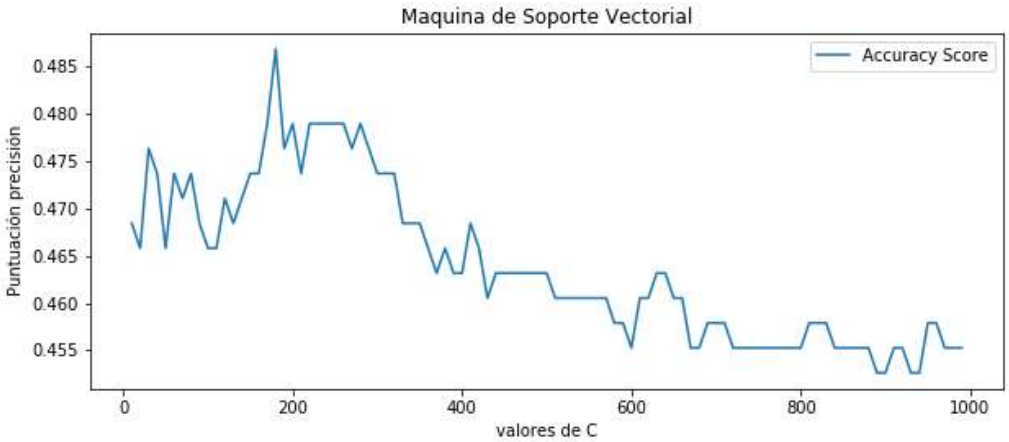
(Arias, 2019)

El algoritmo *Random Forest*, con respecto a regresión logística logra ajustarse un poco mejor a los a la clase 'Empates', lo anterior, de acuerdo con el número establecido de árboles y la profundidad de estos.

2.5.3 Máquina de soporte vectorial. Para implementar el algoritmo de Máquinas de Soporte Vectorial se tuvieron en cuenta varios parámetros en la implementación de la librería *Scikit Learn* de Python, el cual uno de ellos es el parámetro '*C*', es el parámetro de penalización del término de error. Controla la compensación entre el límite de decisión y la clasificación correcta de los puntos de entrenamiento. El otro parámetro es el *Kernel* el cual selecciona el tipo de hiperplano para separar los datos, para el caso de esta implementación se utilizará el *Kernel* 'rbf' (*Radial basis function*), ya que utiliza un hiperplano no lineal que logra adaptarse mejor a los datos.

Para poder encontrar el valor de '*C*' que proporcionara una mayor puntuación de precisión, se procede a realizar una iteración sobre dicha la variable, posteriormente se gráfica y se halla el valor de '*C*' con la mejor puntuación, la variación de la puntuación de precisión en función del parámetro 'C' se observa en la ilustración 43:

Ilustración 43. Exactitud en función de la constante de penalización.

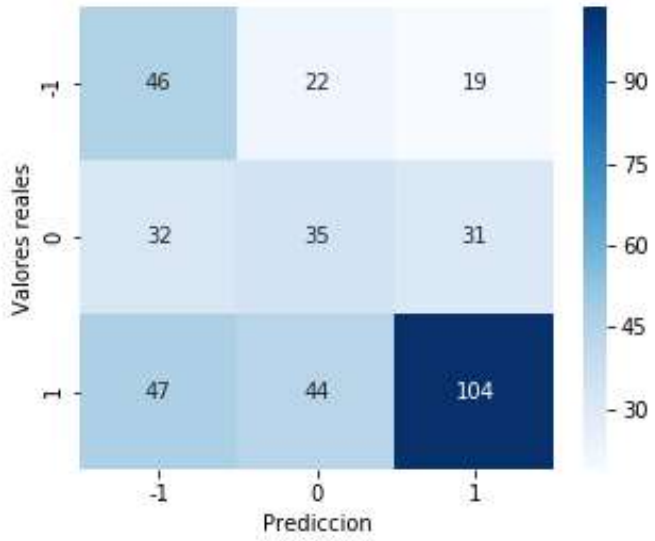


(Arias, 2019)

Una vez ejecutado las iteraciones sobre el parámetro 'C' el valor que presenta una puntuación de precisión más alta es 310. En base a lo anterior, se procede a aplicar

la métrica de precisión matriz de confusión para identificar la clasificación realizada por el algoritmo como se observa en la ilustración 44.

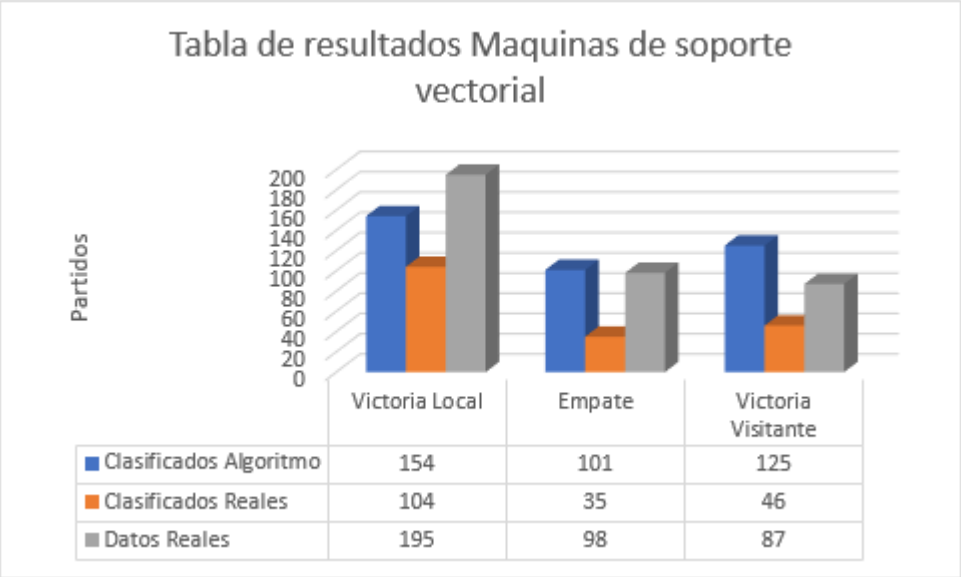
Ilustración 44. Matriz de confusión máquinas de soporte vectorial.



(Arias, 2019)

De acuerdo con la ilustración 44, la matriz de confusión indica que el algoritmo Máquinas de soporte categorizo 154 partidos como victoria de equipos locales, siendo categorizados de forma correcta 104 partidos de 195 partidos reales ganados por equipos locales. Posteriormente el algoritmo categorizó 125 partidos ganados por equipos visitantes, siendo categorizados de forma correcta 46 de 87 partidos reales ganados por equipos visitantes y finalmente 101 partidos categorizados como empate, categorizando de forma correcta 35 de 98 empates reales según el conjunto de datos de prueba, lo anterior se puede visualizar gráficamente en la tabla 5:

Tabla 5. Tabla de resultados máquinas de soporte vectorial.



(Arias, 2019)

Ilustración 45. Reporte de resultados máquinas de soporte vectorial.

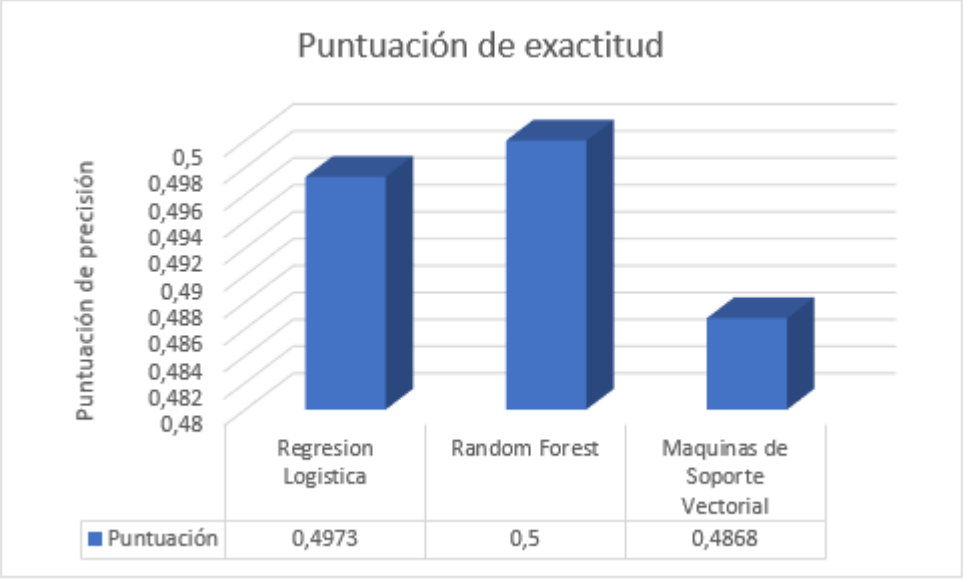
```
accuracy score: 0.5263157894736842

                precision    recall  f1-score   support

Victoria Visitante       0.37      0.53      0.43        87
Empate                   0.35      0.36      0.35        98
Victoria Local           0.68      0.53      0.60       195

    micro avg            0.49      0.49      0.49       380
    macro avg            0.46      0.47      0.46       380
weighted avg             0.52      0.49      0.50       380
```

(Arias, 2019)

A pesar de que el algoritmo máquina de soporte vectorial tiene el mejor *recall* con respecto a los demás algoritmos para los casos en que gana un equipo local, en las demás clases tiene una baja puntuación por lo que no es un algoritmo que prediga de forma eficientemente el ganador de un partido de fútbol de forma equitativa por cada una de las clases, teniendo en cuenta el conjunto de datos del fútbol profesional colombiano.

2.5.4 Puntuación algoritmos de aprendizaje de máquina. La tabla 6 muestra la puntuación de precisión de cada uno de los algoritmos de aprendizaje de máquina para predecir el ganador de un partido de la categoría A del fútbol profesional colombiano en base a las temporadas 2015 a 2018.

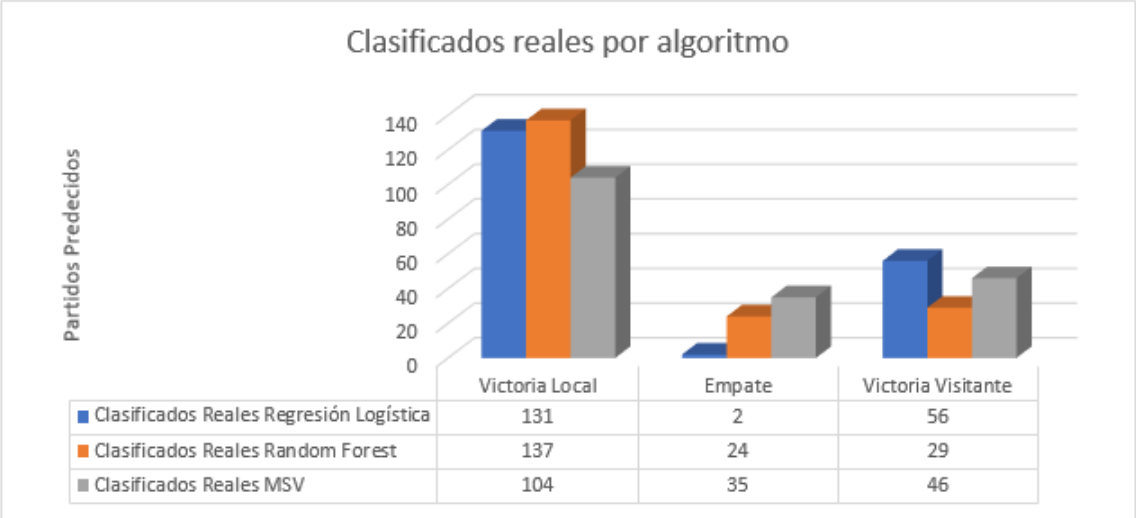
Tabla 6. Tabla general puntuación de precisión.



(Arias, 2019)

En la tabla 7, están consolidado los resultados por cada una de las clases la cuales fueron categorizadas de forma correcta por cada uno de los algoritmos, como se puede ver en la tabla, el algoritmo con mejor predicción es *random forest*, ya que tiene la mayor cantidad de partidos categorizados de forma correcta con 190 de 380 de conjunto de prueba, exactamente la mitad de los datos. Por debajo de *random forest* se encuentra el algoritmo regresión logística con 189 partidos categorizados de forma correcta, con la observación de que los partidos categorizados como empate están muy por debajo de los demás algoritmos. Y finalmente el algoritmo máquina de soporte vectorial el cual el cual tiene la menor cantidad de partidos categorizados de forma correcta, con la observación de que la clasificación en las clases empates y victorias de equipos visitantes es superior a los demás algoritmos.

Tabla 7. Partidos por algoritmo.



(Arias, 2019)

3. CONCLUSIONES

En el presente documento se implementaron 3 algoritmos de clasificación: regresión logística, *random forest* y máquinas de soporte vectorial, lo anterior, en base a la construcción de un conjunto de datos con su respectivo análisis y selección de características asociadas a la variable dependiente, con el objetivo de predecir ganador de un partido de la categoría A del fútbol profesional colombiano. En el transcurso del desarrollo del presente trabajo de grado se identificaron inconvenientes en base a la información del conjunto de datos, ya que las clases no se encontraban balanceadas. Al intentar desarrollar el proyecto sin realizar un balanceo de clases, todos los algoritmos adoptaron un patrón de comportamiento en el cual las predicciones de los datos se centraban en la clase mayoritaria, para este caso, predecía con mayor frecuencia ganadores en condición de visitante.

Al ejecutar los algoritmos de aprendizaje de máquina con el hiperparametro *class_weight* presente en cada uno, los algoritmos adoptaron un comportamiento diferente, pues ya las predicciones no se centraban en la clase mayoritaria producto del desbalanceo de clases del conjunto de datos, se identificó una mejora considerable, ya que fue mayor la predicción de partidos en las clases minoritarias.

A pesar de que los algoritmos tuvieron una mejora notable, la precisión de predicción no subió los esperado, pues ya que los atributos del conjunto de datos son limitados y no cuentan con características relevantes que en efecto puedan impactar en quien será el ganador de un partido de fútbol, como, por ejemplo, datos de tipo climatológico, factores emocionales, puntuación del árbitro, factores sociales y probabilidades de apuestas asociadas a los partidos.

En cuanto a la pregunta de investigación ¿Pueden los algoritmos de aprendizaje de máquina predecir el ganador de un partido de la categoría A del fútbol profesional colombiano, basado en la información de los resultados de los años 2015 a 2018? En efecto, los algoritmos de aprendizaje de máquina pueden predecir el ganador de un partido del futbol profesional colombiano, no obstante, en el presente trabajo de investigación tecnológica como se pudo observar en la sección 2.5.4 las probabilidades de cada uno de los algoritmos de aprendizaje de máquina son cercanas al 0.5, es decir, no existe un criterio claro y definido para predecir el ganador de un partido en función de la probabilidad que se obtuvo como resultado.

4. RECOMENDACIONES Y TRABAJO FUTURO

4.1. RECOMENDACIONES

Se recomienda realizar el experimento con otros algoritmos de aprendizaje de máquina, con el objetivo de evaluar el poder precisión de cada uno y determinar el que mejor se ajuste a los datos reales, por otra parte, se recomienda ampliar el conjunto de datos no en cuanto registros, sino en atributos relevantes que ayuden a mejorar la puntuación de precisión de los algoritmos y así mismo obtener mejores predicciones.

4.2. TRABAJO FUTURO

Los trabajos a futuro a realizar con en base a las dificultades encontradas en el presente trabajo, se encuentra realizar una obtención de variables que permitan obtener una mejor predicción, como, por ejemplo: característica propias de los jugadores que conforman el equipo de fútbol y medir su a rendimiento en diferentes aspectos, factores emocionales dentro del equipo de fútbol, condiciones climáticas y estado del campo de juego, la calidad del arbitraje del partido y así, poder construir un modelo que mejor se ajuste al fútbol profesional colombiano.

BIBLIOGRAFÍA

Anand, G. (2018). English Football Prediction Using Machine Lerning Classifiers. *International Journal of Pure and Applied Mathematics*, 533-535.

Anzola, N. S. (2015). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. *revistas.uexternado.edu.co*.

Arias, E. F. (29 de Abril de 2019).

betegy. (2012). *betegy.com*. Obtenido de betegy.com: https://betegy.com/about/company

betegy.com. (2012). *betegy.com*. Obtenido de betegy.com: https://betegy.com/about/company

Bowles, M. (2015). *Machine Learning in Python.* Indianapolis: Wiley.

Brownlee, J. (25 de Diciembre de 2013). *machinelearningmastery.com.* Obtenido de machinelearningmastery.com: https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

Bunker, R. P., & Thabtah, F. (2017). *sciencedirect.* Obtenido de sciencedirect: https://ac.els-cdn.com/S2210832717301485/1-s2.0-S2210832717301485-main.pdf?_tid=20360d51-bfb6-40e2-8ba6-0ee3e1adb2d9&acdnat=1542065351_a4451c50a515882604fc3b733c53bf0c

code.i-harness. (2016). *CODE Q&A*. Obtenido de CODE Q&A: https://code.i-harness.com/es/q/1bf48c

code.i-harness.com. (2019). *CODE Q&A*. Obtenido de CODE Q&A: https://code.i-harness.com/es/q/15caef4

developers.google.com. (1 de Mayo de 2019). Obtenido de developers.google.com: https://developers.google.com/machine-learning/crash-course/glossary?hl=es-419

dimayor.com.co. (2018). *dimayor.com.co*. Obtenido de dimayor.com.co: http://dimayor.com.co/

dinero.com. (8 de Febrero de 2013). *Revista Dinero*. Obtenido de Revista Dinero: https://www.dinero.com/pais/articulo/sus-derechos-habeas-data/181020

Ekefre, D. B. (14 de 2 de 2016). A Comparison of Methods for Predicting Football Matches. Netherland.

eltiempo.com. (1 de Abril de 2018). *eltiempo.com*. Obtenido de eltiempo.com: https://www.eltiempo.com/colombia/conozca-el-metodo-que-ayuda-a-predecir-resultados-de-futbol-199456

Esme, E., & Kiran, M. S. (2018). Prediction of Football Match Outcomes Based on Bookmaker Odds by Using k-Nearest Neighbor Algorithm. *International Journal of Machine Learning and Computing,* . Obtenido de http://www.ijmlc.org/vol8/658-A05.pdf

FIFA. (1 de Enero de 2006). *FIFA*. Obtenido de FIFA: https://www.fifa.com/mm/document/fifafacts/bcoffsurv/bigcount.statspackage_7024.pdf

FIFA. (2018). *FIFA*. Obtenido de FIFA: https://es.fifa.com/about-fifa/who-we-are/the-game/index.html

Gandhi, R. (13 de Junio de 2017). *towardsdatascience.com.* Obtenido de towardsdatascience.com: https://towardsdatascience.com/k-nearest-neighbours-introduction-to-machine-learning-algorithms-18e7ce3d802a

Gandhi, R. (27 de Mayo de 2018). *towardsdatascience.com*. Obtenido de towardsdatascience.com: https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

Ganesan, A. (2018). ENGLISH FOOTBALL PREDICTION USING MACHINE LEARNING CLASSIFIERS.

Gaviria, A. H. (7 de Diciembre de 2018). *ligadeportiva.com.* Obtenido de ligadeportiva.com: https://ligadeportiva.com/andres-ricaurte-puede-suceder-cualquier-cosa/

GeoTutoriales. (15 de 6 de 2018). *gestiondeoperaciones.net*. Obtenido de gestiondeoperaciones.net: https://www.gestiondeoperaciones.net/proyeccion-de-demanda/calculo-de-la-raiz-del-error-cuadratico-medio-o-rmse-root-mean-squared-error/

Gónzalez, L. (5 de Abril de 2019). *Bosque Aleatorios Regresión – Teoría*. Obtenido de Bosque Aleatorios Regresión – Teoría: http://ligdigonzalez.com/bosques-aleatorios-regresion-teoria-machine-learning/

Hassan, A. R., & Londoño, M. G. (28 de Junio de 2016). *besmarter-team.org.* Obtenido de besmarter-team.org: http://www.besmarter-team.org/files/working_papers/Profiting%20from%20the%20English%20Premier%20League%20Objective%20Predictive%20Elicitation,%20the%20Kelly%20Criterion%20and%20Black%20Swans.pdf

Hijmans, A. (2016). *pdfs.semanticscholar.org.* Obtenido de pdfs.semanticscholar.org: https://pdfs.semanticscholar.org/b347/e38d5c61a139115884fbff352221c4f7bfe1.pdf

ICEMD. (13 de 1 de 2017). *ICEMD*. Obtenido de ICEMD: https://www.icemd.com/digital-knowledge/articulos/mineria-datos-proceso-areas-se-puede-aplica/

Juergen, M. (23 de Mayo de 2013). *entrepreneur.com*. Obtenido de entrepreneur.com: https://www.entrepreneur.com/article/226710

Maher, M. (1982). Modelling association football scores. *Statistica Neerlandica*, 109-118. Obtenido de http://www.90minut.pl/misc/maher.pdf

Marin, L. É. (30 de Junio de 2017). *elcolombiano*. Obtenido de elcolombiano: http://www.elcolombiano.com/deportes/futbol/las-marcas-de-futbol-mas-valiosas-de-colombia-AM6817727

Marr, B. (19 de Febrero de 2016). *Forbes*. Obtenido de Forbes: https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#1a389fb915e7

medium.com. (27 de Diciembre de 2017). *medium.com*. Obtenido de medium.com: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Mehta, A. (9 de Octubre de 2017). *analyticsinsight*. Obtenido de analyticsinsight: https://www.analyticsinsight.net/how-analytics-is-making-teams-win-in-sports/

ml-cheatsheet. (2017). *ml-cheatsheet*. Obtenido de ml-cheatsheet: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#introduction

Momoh, O. (9 de Marzo de 2018). *investopedia.com/*. Obtenido de investopedia.com/: https://www.investopedia.com/terms/d/deep-learning.asp

numberfire.com. (2018). *numberfire.com*. Obtenido de numberfire.com: http://www.numberfire.com

optasports. (2018). *optasports.com*. Obtenido de optasports.com: https://www.optasports.com/

Peréz, D. (22 de Agosto de 2017). *objetivoanalista.* Obtenido de objetivoanalista: https://objetivoanalista.com/big-data-en-el-real-madrid/

Pérez, D. (22 de Agosto de 2017). *ObjetivoAnalista*. Obtenido de ObjetivoAnalista: https://objetivoanalista.com/big-data-en-el-real-madrid/

Rieuf, E. (11 de Febrerp de 2017). *datasciencecentral.com*. Obtenido de datasciencecentral.com: https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the

Soni, D. (2017). *towardsdatascience.com*. Obtenido de towardsdatascience.com: https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54

sportscourtdimensions.com. (2015). *sportscourtdimensions*. Obtenido de sportscourtdimensions: https://www.sportscourtdimensions.com/soccer/

stats.com. (2018). *stats.com*. Obtenido de stats.com: https://www.stats.com/

Varone, M. (2018). *expertsystem.com*. Obtenido de expertsystem.com: https://www.expertsystem.com/machine-learning-definition/

Vincent, J. (6 de Julio de 2017). *the verge*. Obtenido de the verge: https://www.theverge.com/2017/7/6/15923784/ai-predict-sport-betting-gambling-stratagem

ANEXOS

El conjunto de datos, las imágenes utilizadas en este documento y el experimento de aprendizaje de máquina para la predicción del ganador de un partido de la categoría A del fútbol profesional colombiano que se realizó en el presente trabajo de investigación tecnológica, la encontrará en el siguiente repositorio en GitHub:

https://github.com/efarias04/footballPrediction